Functional characterization of genetic variation with in silico predictions of cell-type-specific regulatory elements

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

VE RI TAS

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division of Medical Sciences

in the subject of Biomedical Informatics

have examined a dissertation entitled

Functional characterization of genetic variation with in silico predictions of cell-type-specific regulatory elements

presented by Tiffany Amariuta

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature: ____________________
Typed Name: Dr. Po-Ru Loh

Signature: Zhiping Weng
Zhiping Weng (Aug 26, 2020 16:46 EDT)
Typed Name: Dr. Zhiping Weng

Signature: Alexander Gusev
Alexander Gusev (Aug 26, 2020 21:17 EDT)
Typed Name: Dr. Alexander Gusev

Signature: Eimear Kenny
Eimear Kenny (Aug 26, 2020 15:53 EDT)
Typed Name: Dr. Eimear Kenny

Date: August 25, 2020

Functional characterization of genetic variation with in silico predictions of cell-type-specific regulatory elements

A dissertation presented

by

Tiffany Amariuta

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biomedical Informatics

Harvard University

Cambridge, Massachusetts

August 2020

Dissertation Advisor: Soumya Raychaudhuri Tiffany Amariuta

Functional characterization of genetic variation with *in silico* predictions of cell-type-specific regulatory elements

Abstract

Genome-wide association studies (GWAS) have implicated thousands of complex trait-variant associations, an estimated 90% of which reside in the noncoding genome. While noncoding variants generally have poorly understood regulatory function, previous work has shown that disease-driving genetic variation often affects cell-type-specific gene regulation, such as transcription factor (TF) binding. However, maps of TF-mediated cell-type-specific regulation are currently incomplete due to limited amounts of experimental data. In this thesis, I introduce a novel strategy to annotate the noncoding genome with cell-type-specific regulatory element probabilities via integration and modeling of thousands of publicly available epigenetic datasets. I show that these functional annotations in the disease-driving cell type are more highly enriched for disease heritability than experimentally derived functional annotations. Next, I use these functional annotations to prioritize disease-relevant variants in the context of polygenic risk score (PRS) models. I show that this approach improves the trans-ethnic portability of PRS by reducing the confounding effects of population-specific linkage disequilibrium. Lastly, I introduce a novel strategy to leverage the unprecedented resolution of single cell data to elucidate cell-state-specific activity of trait-driving variants identified by polygenic fine-mapping data from GWAS. This strategy consists of calculating cell-specific enrichments of genome-wide genetic variation in functional regions and then associating these enrichments with polygenic regulatory programs. I show that this approach identifies heterogeneity of risk variant accessibility, nominating putatively causal cell states and regulatory mechanisms. Altogether, this work demonstrates the importance of comprehensive functional annotations to better understand disease and trait etiology.

Table of Contents

Acknowledgments

I would like to thank my advisor Soumya Raychaudhuri for his support, mentorship and guidance during these last five years. I would also like to thank Alkes Price for his continued support, active interest, and collaborative role in supervising my work through the many scientific links between our groups. I would also like to thank Kazuyoshi Ishigaki, Steven Gazal, Bryce van de Geijn, Yang Luo, and Emma Davenport, my main collaborators and co-authors on published work. I am thankful for my family, especially my fiancé Eric Bartell, who has been an inspirational support system for me during my PhD and endured far too many practice presentations and brainstorming sessions.

Chapter 1

Introduction

This thesis outlines the research I have led while advised by Professor Soumya Raychaudhuri at Harvard Medical School toward unraveling the biological mechanisms driving human disease. The genetic code in our DNA predisposes us to different traits and diseases. For many traits and diseases, multiple factors contribute to this predisposition. Knowledge of the biological mechanisms affected by these factors can enhance our understanding of these diseases and ultimately propose hypotheses guiding the development of therapeutic treatments. In this thesis, I will focus on specifically the genetic factors, as opposed to environmental, that regulate human traits and diseases. For a minority of traits and diseases these genetic factors and their biological mechanisms are well understood. However, for human traits and diseases driven by multiple genetic factors, biological explanations for the coordinated and genome-wide roles played by these factors are not well understood. This is largely due to incomplete biological or functional annotation of the majority of the genome, which is precisely the area to which I hope my thesis work has contributed. I will begin with an introduction to the area of human genetics that is relevant to this work in order to evaluate the potentials for advancement of knowledge that I pursued in my thesis work regarding functional characterization of genetic variation.

Complex traits and diseases are a class of phenotypes driven by multiple genetic and environmental factors [1]. These are in contrast to Mendelian traits and diseases that are driven by a single genetic determinant. Human Mendelian traits and diseases were extensively studied during the early years of human genetics predominantly by linkage studies, in which familial inheritance patterns of genetic markers indicating the approximate locations of genes could be traced. Linkage studies revealed large effect genetic determinants of complex traits and diseases, but could not reveal the multiple other smaller genome-wide effects. For this reason, the mechanisms underpinning complex traits and diseases are far less well understood that those of Mendelian traits. For example, early studies of rheumatoid arthritis (RA) using serological typing revealed that human leukocyte antigen (*HLA*) genes within the major histocompatibility complex (MHC) were strongly associated with the disease [2]. This association was later confirmed by familial linkage studies [3]. The ensuing years of research revealed that the MHC alone was not sufficient to explain the genetic variation observed in RA patients, suggesting that other genetic determinants of smaller effect that were missed by linkage studies also contributed.

Linkage studies failed to reveal genetic determinants of smaller effect for several reasons. First, linkage studies often identified large linkage peaks which implicated many genes which were difficult to prioritize. Second, at the time when linkage studies were most prevalent, gene annotations were limited and incomplete compared to the annotation of current day, biasing the associations to well-annotated genes with nearby traceable genetic markers. In order to identify the genetic determinants of complex traits and diseases missed by linkage studies, Botstein and colleagues proposed in 2003 that genome-wide single nucleotide polymorphism (SNP) association studies must be prioritized over linkage studies [4]. These studies provided a framework to test for the association between phenotypes and genotyped SNPs, rather than a limited set of candidate genes as in linkage studies, with the possibility to nominate a causal gene or related regulatory region. SNP-level association studies propelled our understanding of complex traits and diseases. For example, associations with RA were identified in genes beyond the MHC including *PTPN22*, *PADI4*, and *CTLA4* [5–7]. However, the ability to identify these genetic determinants of smaller effect using genome-wide SNP-level association studies greatly depended on the number of individuals in the study. As a consequence, many early association studies reported what were later identified as irreproducible findings due to power limitations.

In the coming years, two foundational projects would lay the groundwork for performing the high-powered, large scale genome-wide association studies (GWAS) that we are familiar with today. First, the completion of the Human Genome Project in 2001 revealed for the first time 94% of the base pairs in the human genome and 1.4 million SNPs [8]. Performing whole genome sequencing on so many variants with the large sample size required to confidently identify associations with complex traits and diseases would be prohibitively expensive. However, there is substantial correlational structure of variants across the genome and this could be leveraged to reduce costs. Regions of the genome that undergo less recombination are inherited together, resulting in blocks of SNPs with highly correlated genotypes. This phenomenon is called linkage disequilibrium (LD). To mitigate the costs of GWAS, the International HapMap Project then proposed designing a genotyping chip with a subset of representative SNPs, each marking a different LD block. Identified associated variants would point to a candidate causal locus, and further statistical or functional work would need to be done to identify the true causal SNP in the locus. This enabled an era of high-powered GWAS in which thousands of individuals could be genotyped at reasonable cost, exponentially adding to the list of putatively causal genetic determinants of complex traits and diseases.

Although the number of genome-wide significant associations were quickly increasing for complex traits and diseases, the understanding of the mechanisms through which these variants act was only slowly advancing. This is because the noncoding genome harbors an estimated 90% of genome-wide significant variants. For the approximately 10% of associated variants that reside in coding regions, it is more straightforward to hypothesize mechanisms involving the implicated gene. Noncoding variation is challenging to understand because the noncoding genome is less well annotated, for example, with enhancers, promoters, and important regulatory elements and their cell-type-specific counterparts, than the coding genome, with genes. Understanding the mechanisms of noncoding variation is crucial to understanding the biology underlying complex traits and diseases. More than a decade ago, the field of human genetics began producing strategies to link noncoding variants to functional biology, as outlined below.

Most strategies to mechanistically link noncoding variants to complex traits and diseases try to understand the effect of the noncoding variation on gene expression. Colocalization studies of genome-wide significant noncoding variants with expression quantitative trait loci (eQTLs) nominated novel candidate causal genes. For example, studies of RA found novel importance in *CCR6* [9], *AOAH* [10], *BLK*, *C5orf30*, *GSDMB*, *IRF5*, and *PLEK* [11]. However, simple colocalization does not imply the same causal genetic driver. In 2017, Chun and colleagues devised a strategy to test if the GWAS association signal and eQTL association signal were produced by the same genetic determinant [12]. They found that a discouragingly small proportion (~25%) of noncoding variation can be attributed to modulating gene expression levels, as measured in an eQTL study.

While colocalization studies considered biological mechanisms of noncoding variants from the perspective of specific genes, more recent studies considered biological mechanisms from the perspective of the cell type specificity of gene expression programs. For example, the genes identified by eQTL colocalization studies of RA had implicated many different immune cell types, but there was no quantitative understanding of the relative contributions of these cell types. Moreover, eQTLs are more likely to be cell-type-nonspecific, as we are less powered to identify cell-type-specific signals. Work from our group hypothesized that genetic risk factors for non-systemic diseases and traits act via mechanisms that affect a small set of tissues or cell types [13]. In this study, Hu and colleagues assessed the enrichment of cell-type-specific gene expression in RA risk loci, revealing CD4+ effector memory T cells as the strongest candidate causal cell type.

Studies based on gene expression were inherently limited to noncoding variation that could be associated with genes, which in the case of eQTL colocalization preferentially selected noncoding variation proximal to the gene. If genes with cell-type-specific expression profiles were enriched for complex trait and disease risk loci, then cell-type-specific gene regulatory elements should as well. The advantage to this perspective was that noncoding genetic variation could be more comprehensively studied. As a result, many studies assessed the colocalization of cell-type-specific epigenetic marks with complex trait and disease risk loci. A study from our group identified the strongest colocalization of RA risk loci with H3K4me3 from CD4+ regulatory T cells [14]. Another study investigated the colocalization of risk loci with DNase hypersensitivity sites (DHSs), which they correlated with gene expression to ultimately link the risk locus to a putative target gene [15]. A third study coupled statistical fine-mapping, a strategy to deconvolute the correlation between variant associations due to LD and identify the most likely causal variant, with epigenetic colocalization [16]. Considering 21 autoimmune diseases, this study found that 60% of fine-mapped putatively causal variants colocalized with CD4+ T cell enhancers.

While epigenetic marks may colocalize even with each other, it is of interest to know which regulatory annotation distinguishes best between causal and non-causal variants. For example, transcription factors (TFs) often bind in cell-type-specific manners. If a set of disease risk loci colocalize with ChIP-seq peaks of a particular TF, this might implicate the TF in a causal disease-driving mechanism. However, TF ChIP-seq peaks often colocalize with gene promoters and enhancers, as TFs are precisely recruited there to modulate gene expression. Therefore, colocalization of risk loci with TF ChIP-seq peaks might be the result of unaccounted and stronger colocalization with gene promoters in general. Thus, we must ask if the enrichment of risk variants in TF ChIP-seq peaks is still significant once conditioning on the promoter association. Prior work from our group addressed this question with a method called Genomic Annotation Shifter (GoShifter) [17]. This approach statistically quantifies the enrichment of risk loci in regulatory annotations via permutation, while explicitly controlling for two sources of confounding bias: 1) LD and 2) coincidental colocalization of the query regulatory annotation with an annotation that better distinguishes casual from non-causal variants.

Thus far, the discussed strategies to link noncoding genetic variation to functional mechanisms have relied on the identification of genome-wide significant variants identified by GWAS. For many complex traits and diseases, there are too few genome-wide significant variants to perform the aforementioned strategies. For example, GWAS of schizophrenia historically reveal few genome-wide significant variants at best. A previous GWAS, with over 6,000 individuals, identified no genome-wide significant variants [18]. For complex traits and diseases driven by many small genetic effects across the genome, GWAS with current sample sizes are not powered to confidently identify these associations. While the total genetic variance, or heritability, for many traits and diseases had been estimated in familial studies in past years, the proportion of total heritability explained by genome-wide significant variants turned out to be discouragingly low. This phenomenon was coined as the problem of “missing heritability”. While there are many possible explanations for this, including 1) causal variants of small effect size cannot be identified with current GWAS sample sizes, 2) heritability quantified in familial studies may be overestimated, and 3) genotyping chips exclude low frequency variants, rare variants, and copy number variants. The same study that turned up no genome-wide significant variants in a schizophrenia GWAS estimated that at least one-third of the liability of the disease is attributable to common polygenic variation, undetected by GWAS. Moreover, the authors demonstrated that schizophrenia cases had higher genetic risk scores, an aggregate score of one's genotype weighted by variant effect size, than controls.

These challenges in studying complex traits and diseases led to strategies that model association statistics from common variants irrespective of their genome-wide significance. In RA genetics, one such strategy took a similar approach to the previously discussed schizophrenia GWAS study. In 2012, Stahl and colleagues used a polygenic risk score approach to attribute 20% of RA heritability to 2.5 million common variants, a contribution independent of the 25% of RA heritability attributed to the MHC [19]. Around the same time, Yang and colleagues devised an approach to estimate the total SNP heritability of complex traits and diseases by computing the association between groups of variants and phenotypes, as opposed to the traditional GWAS approach which finds associations between single SNPs and phenotypes [20]. This approach, widely referred to as GCTA, established a new gold standard for total SNP heritability estimates, replacing estimates from family studies. With GCTA, not only was it possible to estimate the total SNP heritability of a trait, but also the relative contribution of categories of SNPs, for example, implicating functional or regulatory programs. In 2012, Lee and colleagues found that specifically expressed genes in the central nervous system were disproportionately enriched for schizophrenia heritability [21]. Then Gusev and colleagues found that variants residing in cell-type-specific regulatory elements were strongly enriched for heritability of a variety of complex traits [22]. These studies began to not only identify but quantify the contribution of different biological processes underpinning complex traits and diseases.

Soon thereafter, Bulik-Sullivan and colleagues devised a faster and more accessible approach called LD score regression (LDSC) to quantify the total common SNP heritability of complex traits and diseases [23]. Concurrently, Finucane and colleagues developed a derivative approach called stratified LDSC (S-LDSC) to quantify the contribution of different categories of variants to that total SNP heritability [24]. These approaches revolutionized the study of complex traits and diseases by overcoming many of the shortcomings of previous methods. First, LDSC and derivative methods considered all SNPs genome-wide irrespective of GWAS association, but usually enforcing a minor allele frequency (MAF) lower bound. Second, these methods did not assume one causal variant per locus as did strategies considering only genome-wide significant variants and further restricting to the lead association or lead fine-mapped variant in the locus. Third, these methods utilized summary association statistics from the GWAS and did not require individual-level genotyping data, which often do not exist in the public domain. In this thesis, we will use S-LDSC as the state-of-the-art approach to partition the heritability of complex traits and diseases by functional category of SNPs.

Studies that partitioned the common SNP heritability of complex traits and diseases with functional annotations indicating cell-type-specific histone marks[24], cell-type-specifically expressed gene sets[25], and directional effects of TF binding[26] shed light on the biological mechanisms of coordinated genetic variation. However, the functional annotations of these studies are limited in their specificity to disease-relevant biological processes. For example, presence of the H3K4me3 histone mark indicating active promoters and enhancers in CD4+ regulatory T cells might be enriched for RA heritability, but do not mark regions specific to CD4+ regulatory T cells. These active promoters and enhancers comprise two categories: 1) those associated with cell-type-nonspecific cell cycle and housekeeping processes and 2) those associated with the difference in lineage specification of CD4+ regulatory T cells from other memory T cells. Therefore, we hypothesized that the functional annotation that best captures disease heritability would be the one implicating pathogenic and cell-type-specific activity within the disease-driving cell type. In order to focus on the regulatory elements that confer pathogenic identity to disease-driving cell types, we focus on the targets of master regulator TFs that specify differentiation paths of naive cell types to mature lineages. We further aimed to create functional annotations that could prioritize risk variants, and thus would consist of probabilistic SNP-level scores, as opposed to binary membership to the functional category as in previous studies[24,25]. We also aimed to create functional annotations that would aggregate *in silico* via predictive modeling thousands of experimental datasets to produce a comprehensive track of disease-relevant cell-type-specific regulatory activity, as opposed to the common use of individual experimental datasets, susceptible to noise and variation.

In this thesis, I describe the many genetic and genomic applications of designing functional annotations that better capture both disease-relevant and cell-type-specifying regulatory elements.

In Chapter 2, I describe our strategy to identify regulatory elements that capture substantially larger proportions of complex trait and disease heritability than commonly used functional annotations and the generalizability of our approach to any complex trait or disease. Briefly, we utilize genome-wide protein occupancy profiles of master regulator transcription factors as a basis to learn an epigenetic signature that might be representative of all cell-type-defining regulatory elements. We demonstrate the validity of this approach by the improvement of captured polygenic heritability compared to widely used functional annotations derived from experiments, including histone modification ChIP-seq and RNA-sequencing.

In Chapter 3, I describe our use of these functional annotations to reduce confounding bias in genetic association data to improve multi-ethnic transferability and study shared regulatory mechanisms. When quantifying the contributions of cell-type-specific regulatory mechanisms, modeled by IMPACT, to a diverse set of complex traits and diseases, we identified an overwhelmingly strong concordance between European and East Asian populations. We then found that prioritizing variants in predicted disease-driving cell-type-specific regulatory elements improved the predictive accuracy of PRS models built using European genetic data and applied to an East Asian population.

In Chapter 4, I describe our approach to leverage single cell epigenetic data in order to identify cellular subpopulations with different pathogenic potentials. Specifically, we perform multi-modal data integration involving polygenic fine-mapping, single cell chromatin accessibility assays, and functional annotation data. Our analysis revealed potential trait-driving regulatory mechanisms of identified cellular subpopulations.

Finally, in Chapter 5, I discuss the broader implications and limitations of this work as a whole as well as potential future directions.

Chapter 2

IMPACT: Genome-wide annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors

The material in this chapter appeared in the May 2019 edition of the *American Journal of Human Genetics* as “IMPACT: Genome-wide annotation of cell-state-specific regulatory elements inferred from the epigenome of bound transcription factors” by Amariuta et al[27].

Abstract

Despite significant progress in annotating the genome with experimental methods, much of the regulatory noncoding genome remains poorly defined. Here we assert that regulatory elements may be characterized by leveraging local epigenomic signatures where specific transcription factors (TFs) are bound. To link these two features, we introduce IMPACT, a genome annotation strategy which identifies regulatory elements defined by cell-state-specific TF binding profiles, learned from 515 chromatin and sequence annotations. We validate IMPACT using multiple compelling applications. First, IMPACT distinguishes between bound and unbound TF motif sites with high accuracy (average AUPRC 0.81, s.e. 0.07; across 8 tested TFs) and outperforms state-of-the-art TF binding prediction methods, MocapG, MocapS, and Virtual ChIP-seq. Second, in eight tested cell types, RNA polymerase II IMPACT annotations capture more cis-eQTL variation than sequence-based annotations, such as promoters and TSS windows (25% average increase in enrichment). Third, integration with rheumatoid arthritis (RA) summary statistics from European (N=38,242) and East Asian (N=22,515) populations revealed that the top 5% of CD4+ Treg IMPACT regulatory elements capture 85.7% of RA h2, the most comprehensive explanation for RA h2 to date. In comparison, the average RA h2 captured by compared CD4+ T histone marks is 42.3% and by CD4+ T specifically expressed gene sets is 36.4%. Lastly, we find that IMPACT may be used in many different cell types to identify complex trait associated regulatory elements.

Introduction

Transcriptional regulation is the foundation for many complex biological phenotypes, from gene expression to disease susceptibility. However, the complexity of gene regulation, controlled by more than 1,600 human transcription factors (TFs)[28] influencing some 20,000 protein coding genes, has made functional annotation of the regulome difficult. Tens of thousands of genomic annotations have been experimentally generated, enabling the success of unsupervised methods such as chromHMM[29] and Segway[30] to identify global chromatin patterns that better characterize genomic function. However, linking specific regulatory processes to these identified patterns is challenging. Furthermore, although genome-wide association studies (GWAS) have identified ~10,000 trait associated variants across hundreds of polygenic traits[31], most variants lie in noncoding regulatory regions with uncertain function.

With continually increasing numbers of genomic annotations generated from high-throughput experimental assays, *in-silico* functional characterization of variants has growing potential. These assays include genome-wide open chromatin, histone mark, and RNA expression profiling, each separately possible at the single cell level. Initially contributed by genomic consortia, such as ENCODE[32] and Roadmap[33], these assays have become more common place as easy-to-implement protocols have been developed, thereby contributing to the growing rate of genomic annotation generation.

Recently, integration of datasets, particularly those indicating regulatory elements, with GWAS data has successfully led to the identification of categories of disease-driving variants enriched for genetic heritability (h2)[24,25,34]. Such regulatory annotations identify active promoters and enhancers through open chromatin or histone mark occupancy assays in a cell type of interest[14–16,24,25]. However, these annotations include both cell-type-specific and nonspecific elements, the latter of which may affect a wide range of cellular functions that are not necessarily intrinsic to disease-driving cell-states. Therefore, we hypothesized that the identification of regulatory elements specifically driving functional states would help us to not only better characterize regulatory elements genome-wide, but also better capture polygenic h2 of complex traits and diseases. Once the most enriched classes of regulatory elements are recognized, then it may become possible to generate biologically-founded mechanistic hypotheses.

Here, we introduce IMPACT (Inference and Modeling of Phenotype-related ACtive Transcription), a diversely applicable genome annotation strategy to predict cell-state-specific regulatory elements. We take a two-step approach to define IMPACT regulatory elements. First, we choose a single key TF, known to regulate a cell-state-specific process, and then identify binding motif sites genome-wide, distinguishing between those that are bound and unbound using genomic occupancy identified by ChIP-seq in the corresponding cell-state. Here, the term "cell-state-specific" refers to the observed experimental binding sites of a key TF, which itself may not be entirely cell-state-specific, assayed in the target cell-state. Second, IMPACT predicts TF occupancy at binding motif sites by aggregating and performing feature selection on 503 cell-type-specific epigenomic features and 12 sequence features in an elastic net logistic regression model. The IMPACT model framework can easily be expanded to accommodate thousands of epigenomic annotations and is amenable to increasing rates of data generation. From this regression we learn a TF binding chromatin profile, which IMPACT uses to probabilistically annotate the genome at nucleotide-resolution. We refer to high scoring regions as cell-state-specific regulatory elements (**Figure 2-1**). With this approach, we aim to better pinpoint sites of causal variation of gene expression and polygenic trait heritability by modeling trait-driving cell-state-specific regulatory processes.

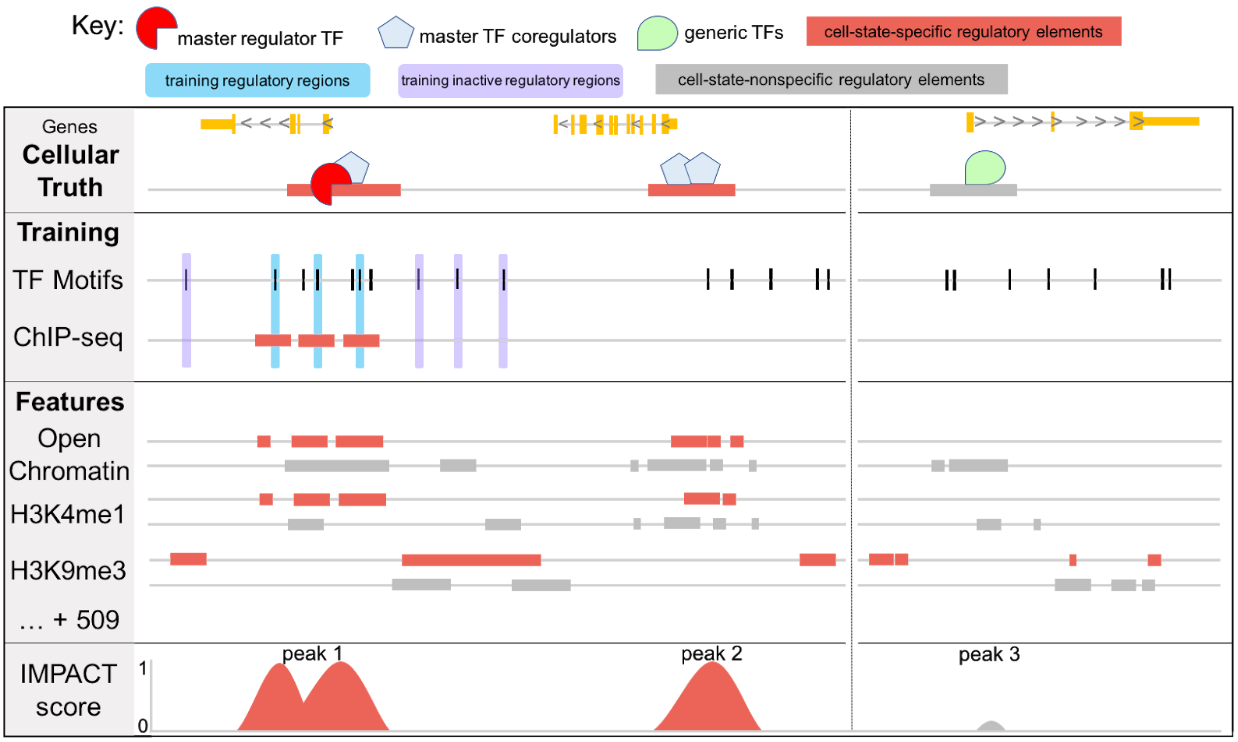


Figure 2-1. IMPACT: a genome annotation strategy to identify cell-state-specific regulatory elements. IMPACT learns a chromatin profile of cell-state-specific regulation, distinguishing master TF (red) regulatory elements (TF-bound motif sites, blue) from inactive regulatory elements (unbound motif sites, purple). Here, cell-state-specific open chromatin and cell-state-specific H3K4me1 are strong predictors of cell-state-specific regulatory elements. Cell-state-nonspecific open chromatin and nonspecific H3K4me1 are less informative, marking all types of regulatory elements, while H3K9me3 strongly implicates inactive regulatory elements. IMPACT should re-identify regulatory elements marked by master TF binding (peak 1) and those with similar chromatin profiles, presumably sites of related cell-state-specific processes (peak 2). IMPACT should not predict regulation at cell-state-nonspecific elements (peak 3), such as promoters of housekeeping genes.

Material and Methods

Data

Genome-wide Annotation Data. We obtained publicly available genome-wide epigenomic annotations including ATAC-seq, DNase-seq, FAIRE-seq, HiChIP, polymerase and elongation factor ChIP-seq, and histone modification ChIP-seq assayed in hematopoietic, adrenal, brain, cardiovascular, gastrointestinal, skeletal, and other cell types for the GRCh37 (hg19) assembly (**Table A-1**). Sequence annotations, downloaded from UCSC, include Phastcons conservation, exons, introns, intergenic regions, 3'UTR (untranslated region), 5'UTR, promoter-TSS (transcription start site), TTS (transcription termination site), and CpG islands. For benchmarking IMPACT against MocapG[35], MocapS[35], and Virtual ChIP-seq[36], we additionally acquired corresponding cell-type-specific open chromatin and gene expression where applicable (**Table A-3**). For models trained on Pol II ChIP-seq, we removed Pol II and elongation factor ChIP-seq feature tracks from the feature library before running IMPACT.

TF ChIP-seq data. We determined genome-wide TF occupancy from publicly available ChIP-seq (**Table A-4**) of 13 key regulators (T-BET[37,38], GATA3[39], STAT3[40], FOXP3[41], STAT5[41], IRF5[42], IRF1[43], CEBPB[44], PAX5[45], REST[32], RXRA[32], HNF4A[46], TCF7L2[32]) assayed in primary cell-states which they have been observed to regulate: Th1, Th2, Th17, Tregs, Tregs, macrophages, monocytes, monocytes, B cells, fetal brain cells, brain cells, liver cells, and pancreatic cells, respectively. We additionally acquired ChIP-seq of RNA polymerase II in peripheral blood/lymphocytes, fibroblasts, stomach, liver, left ventricle heart, sigmoid colon, pancreas, and CD4+ T cells[32,38,47]. All ChIP-seq peaks were called by macs[48] [v1.4.2 20120305] (all P < 1e-5).

cis eQTL data. We acquired SNP-level summary statistics from three independent studies. First, we obtained data from 3,754 peripheral blood samples[49] in which 7,025 unique genes had measurements. As some genes were represented by several array probes, we retained only summary statistics on one probe, selected randomly, per gene. Second, we obtained data from GTEx V7 (**Web Resources**) in the following 6 cell types with the number of samples listed in parentheses: transformed fibroblasts (300), stomach (237), liver (153), left ventricle of heart (272), sigmoid colon (203), and pancreas (220). On average across these cell types, approximately 22,000 genes had measurements in the GTEx data. Third, we obtained eQTL data from CD4+ T cells in East Asian individuals (N=103) with expression measurements for 20,107 genes[50]. For each gene, we truncated the genome-wide summary statistics to a cis window of 1 Mb upstream and downstream of the gene TSS.

Genome-wide association data used in S-LDSC analyses. We collected RA GWAS summary statistics[51] for 38,242 European individuals, combined cases and controls, and 22,515 East Asian individuals, comprised of 4,873 RA cases and 17,642 controls[52]. We estimated total genome-wide polygenic RA h2 to be about 18% for EUR and 21% for EAS. We further collected 41 other complex trait summary statistics[34,53,54]. Reference SNPs, used to estimate European LD scores, were the set of 9,997,231 SNPs with minor allele count greater or equal than five in a set of 659 European samples from phase 3 of 1000 Genomes Projects[55]. The regression coefficients were estimated using 1,125,060 HapMap3 SNPs and heritability was partitioned for the 5,961,159 reference SNPs with MAF ≥ 0.05. Reference SNPs, used to estimate East Asian LD scores, were the set of 8,768,561 SNPs with minor allele count greater or equal than five in a set of 105 East Asian samples from phase 3 of 1000 Genomes Projects[55]. The regression coefficients were estimated using 1,026,051 HapMap3 SNPs and heritability was partitioned for the 5,469,053 reference SNPs with MAF ≥ 0.05. Frequency and weight files (1000G EUR phase3, 1000G EAS phase3) are publicly available and may be found in our **Web Resources**.

Fine-mapped RA causal variation. Previous work from our group aimed to define the most likely causal RA variant for each locus harboring a genome-wide significant variant[56], identified by a GWAS of 11,475 European RA cases and 15,870 controls[57]. To this end, causal posterior probabilities were computed with the approximate Bayesian factor (ABF), assuming one causal variant per locus. The posteriors were defined as:

$$P_i = ABF_i / \sum_{k=0}^{n} ABF_k \ ,$$

where *i* is the i^{th} variant and *n* is the total number of variants in the locus. As such, the ABF over all variants in a locus sum to 1.

Statistical Methods

IMPACT Model. We build a model that predicts TF binding on a motif site by learning the epigenomic profiles of the TF binding sites. We use logistic regression to model the log odds of TF binding on a motif site, or putative binding site, based on a linear combination of the effects β_j of the *j* epigenomic or sequence features (**Table A-1**), where β_0 is an intercept:

$$log\ (\tfrac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j \ ,$$

where X_j is a value defining some relationship between feature *j* and the motif site and *p* is the probability of TF binding at the motif site. From the log odds, which ranges from negative to positive infinity, we compute the probability of TF binding, ranging from 0 to 1:

$$p = \frac{1}{1+exp(-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_j X_j))} .$$

We use a logistic regression framework with elastic net regularization implemented by the *cv.glmnet* R [v1.0.143] package[58], in which optimal β are fit according to the following objective function,

$$argmin_{\beta} = (||Y - X\beta||^2 + \frac{1}{2}(1 - \alpha)||\beta||^2 + \alpha\, ||\beta||).$$

where Y represents the binary vector indicating TF bound or unbound motif sites, X is a matrix defining the feature characterization of each motif site, and α is the mix term between the ridge (L2), $||\beta||^2$, and lasso (L1), $||\beta||$, penalties, where $0 \leq \alpha \leq 1$. We find that no α significantly outperforms the others (**Figure A-1**). Therefore, we select α = 0.5 to make a compromise between sparsity and information content; enforcing sparsity with lasso performs feature selection thereby helping to avoid overfitting. However, excessive feature selection may remove important information. We use elastic net regularization for two reasons: 1) our model has a large number of features (N > 500) which may result in overfitting if feature selection is not performed (L1 penalty) and 2) the L2 penalty makes the objective function convex, with one stable solution.

Training IMPACT. For each cell-state that we model, we train IMPACT to distinguish cell-state-specific regulatory from non-specific or inactive regulatory regions based on cell-state-specific binding of a single key TF. For training, we define the cell-state-specific regulatory class as TF-bound motif sites and the non-specific or inactive regulatory class as unbound motif sites. To define TF-bound motif sites, we use HOMER[59] [v4.8.3] to scan TF ChIP-seq peaks for *k*-mers with a sequence similarity score, computed from the PWM (position weight matrix) across *k* nucleotides, that is greater than or equal to the TF binding motif detection threshold, empirically determined by HOMER. Specifically, this log-odds detection threshold is equal to the maximum achievable log odds (computed from the PWM) minus an empirically derived acceptable degree of mismatches. A detailed description of this calculation may be found in the HOMER documentation (**Web Resources**). We have observed that at most 3 well tolerated nucleotide mismatches are permitted for every 10 nucleotides. HOMER then scans the genome to assess if a putative motif site exceeds the detection threshold. The motif-specific detection threshold for each TF used in this study can be found in **Table A-2**. To test how sensitive our selection of training data and genomic annotation is to this parameter, we iterated over multiple motif detection thresholds ranging from lenient to strict (**Figure A-2**). We observe that small changes in the motif log-odds detection threshold lead to modest changes in the proportion of peaks with a detectable motif. For example, decreasing the threshold by 0.5, leads to an increase of at most 10% of ChIP-seq peaks with a detectable motif and increasing the threshold by 0.5, leads to a decrease of at most 12% of ChIP-seq peaks with a detectable motif. Regarding genomic annotation, we used IFN-G, the quintessential target gene of the TF T-BET, to demonstrate how IMPACT regulatory element probabilities changed in this locus as a result of changing the motif detection threshold (**Figure A-2**). For the following thresholds 4, 5, 6, 6.2 (0.5 lower than the default T-BET detection threshold), 6.7 (default threshold), 7.2 (0.5

greater than the default threshold), 8, and 9, we find that IMPACT regulatory element probabilities do not significantly vary over the IFN-G locus, suggesting that IMPACT genomic annotation is not sensitive to the motif detection threshold parameter.

For each ChIP-seq peak with at least one motif match, we retain only the coordinates of the highest scoring motif match to use in our training set. This ensures that each instance of a bound motif site is in a separate ChIP-seq peak, which avoids double counting ChIP-seq peaks. In terms of training a logistic regression model, this helps to ensure that no two motif instances are in overlapping or proximal genomic coordinates and may be considered independent. We randomly select 1,000 TF-bound motif sites in each training instance.

To define unbound motif sites, we use the genome-wide TF motif scan performed above and select motif matches that do not overlap the corresponding TF ChIP-seq peaks. Then, we retain the genomic coordinates of these matches. We randomly select 10,000 unbound motif sites in each training instance. We select 1,000 bound motif sites and 10,000 unbound motif sites for the following two reasons. First, of all tested TFs, the smallest dataset contained just over 1,000 bound motif sites. Therefore, to uniformly train IMPACT models across TFs, we required the same number of bound motif sites be used in each instance. Second, for the purpose of genome-wide regulatory annotation, we attempt to make our training data represent hypothesized genome-wide regulatory proportions. To this end, we arbitrarily required 10 times as many unbound motif sites as bound motif sites to reflect an approximate genome-wide ratio of non-regulatory to regulatory elements, respectively[60,61]. For the purposes of benchmarking IMPACT against state-of-the-art methods, we assessed each model's performance on the same sets of motif sites.

IMPACT is trained to distinguish TF-bound motif sites from unbound motif sites by their epigenomic and sequence feature characterization. We build a feature matrix by reporting overlap of an annotation and a motif site with a value of 1 and no overlap with a value of 0. Each feature characterization is represented twice in the model, first with respect to local regions, and second with respect to distal regions. In the local case, for each motif site and for each feature, we quantify direct positional overlap. In the distal case, we quantify feature overlap with a distal nucleotide relative to the motif site. We reason that although a motif site may not directly overlap a particular feature, such as a promoter, it may be informative to know that there is one nearby. For example, we might look 1,000 nucleotides away from the motif site and report feature overlap at either the upstream or downstream position with a single value of 1 or overlap at neither with a value of 0. After parameter optimization, we set this distance value to 1,000 nucleotides (**Figure A-1**). We do not use absolute distance between annotation and motif site to characterize our feature space in the interest of computational efficiency with specific regard to nucleotide-based genome-wide annotation. Furthermore, IMPACT prediction performance for no TF is significantly improved by using the absolute distance feature characterization strategy (all $P > 0.60$) (**Figure A-3**).

We note that using motif site-centric gold standards has multiple advantages over predicting TF binding on entire ChIP-seq peak regions. First, using motif sites serves as a quality control for pioneer TF ChIP-seq data, in which case we know the TF is interacting directly with the DNA. Second, it provides an intuitive interpretation for binary labeling as a motif site may be either bound or unbound. Such binary interpretation is not applicable to ChIP-seq peaks which can each implicate hundreds of nucleotides. Rather than a TF binding uniformly throughout the peak, it is more likely that the ChIP-seq signal is coming from a smaller region of TF binding within the peak, making the use of motif sites an attractive strategy to better localize the signal. Third, it provides TF-specificity by focusing on sequences within the peak that only the TF of interest may interact with, whereas within the coordinates of one ChIP-seq peak multiple TFs may be binding. We also observed that on average IMPACT predicts TF binding significantly better when using motif site-centric gold standards according to the AUPRC performance metric (0.18 average increase in AUPRC; all student's t-test P < 8.3e-36, except for TCF7L2) (**Figure A-3**). Moreover, we find that IMPACT regulatory element probabilities are significantly higher (all P < 0.05, student's t-test) at nucleotides located in both a motif site and a ChIP-seq peak (**Figure A-3**), suggesting that motif sites provide a non-redundant layer of regulatory information beyond ChIP-seq peak signal. These results suggest that IMPACT's ability to score motif sites with higher regulatory potential might be used as a strategy to perform quality control on ChIP-seq peaks.

To train the elastic net logistic regression model, we partition the sets of TF-bound and unbound motif sites by randomly sampling 80% of each set, to be used for 10-fold cross validation (CV), in which these subsets are further partitioned into 90% for training and 10% for testing. The remaining 20%, completely unseen by the CV and not overlapping with the initial 80%, is used as a validation set. In this binary classification problem, probabilistic outputs from the logistic regression are made binary by applying thresholds in the CV. The threshold vector is a sequence from 0 to 1, with resolution of 0.0025, resulting in 401 applied thresholds. We applied IMPACT genome-wide to assign nucleotide-resolution cell-state-specific regulatory element probabilities, using the model learned from the elastic net logistic regression CV.

Interpreting IMPACT regulatory element probabilities. Genome-wide, IMPACT evaluates each nucleotide's regulatory element potential with respect to a particular TF/cell-state pair and assigns a probability to each nucleotide. In order to understand these probabilities, we compare their distribution across the bound motif site class and the unbound class. As expected, we observe significantly higher IMPACT predictions at TF-bound motif sites compared to unbound motif sites (all P < 1e-3, student's t-test); unbound motif sites have regulatory probabilities near 0 (**Figure A-4**). This separation informs the interpretation of genome-wide predictions: truly inactive/non-specific regulatory elements are expected to have predicted values close to 0, rather than an arbitrary or uninterpretable non-zero decision boundary.

cis eQTL causal variation enrichment. We computed a genome-wide enrichment of cis eQTL causal association across various functional annotations. To this end, we gathered gene-based cis-window summary statistics. Then, for each gene and for each annotation, an enrichment was calculated explicitly as:

$$Enrichment_{g,a} = \frac{\left(\sum_{i=1}^{N} \chi^2 g\right) / N}{\left(\sum_{j=1}^{M} \chi^2 g\right) / M},$$

where g is the gene, a is the annotation, N is the number of variants within annotation a, M is the number of variants outside annotation a, i is the i^{th} variant, j is the j^{th} variant, and χ^2 is the chi-squared statistic of the association between gene g and SNP i or j. We then computed genome-wide standard errors by block jackknifing the genome into 200 adjacent bins and computed a distribution of enrichment values when leaving one bin out at a time[24]. This strategy is designed to prevent the genes of any one region of the genome from dominating the enrichment statistic. Furthermore, we used a permutation strategy to establish a null distribution. To this end, we randomly permuted the chi-squared associations in the cis-window of each gene 1,000 times, while matching on 50 LD bins across the cis-window, and recomputed the enrichment with each of the functional annotations. We estimated enrichment significance based on how extreme our result was compared to the permutation distributions.

Partitioning heritability with S-LDSC. We apply S-LDSC[24] (stratified linkage disequilibrium (LD) score regression) [v1.0.0], a method developed to partition polygenic trait heritability by one or more functional annotations, to quantify the contribution of IMPACT cell-state-specific regulatory annotations to 42 complex traits. We annotate common SNPs (MAF ≥ 0.05) with regulatory element probabilities based on cell-state-specific IMPACT models. Then, we run S-LDSC once on the annotated SNPs to compute population-specific LD scores and again to quantify the complex trait heritability captured by our IMPACT annotations. Here, the two statistics we use to evaluate how well our annotations capture causal variation are enrichment and standardized effect size (τ *).

If a_{cj} is the value of annotation *c* for SNP *j*, we assume the variance of the effect size of SNP *j* depends linearly on the contribution of each annotation *c*:

$$Var(\beta_j) = \sum_c a_{cj}\tau_c .$$

where τ_c is the per-SNP contribution from one unit of the annotation a_c to heritability. To estimate τ_c, S-LDSC estimates the marginal effect size of SNP *j* in the sample from the chi-squared GWAS statistic χ_j^2:

$$\chi_j^2 = N\,\hat{\beta}_j^2$$

Considering the expectation of χ_j^2 and following the derivation from Gazal et al 2017[34],

$$[\chi_j^2] = N\,\sum_c(\tau_c\sum_k(a_c(k)r_{jk}^2)\ +\ 1,$$

$$E[\chi_j^2] = N\,\sum_c\tau_c l(j,c)\ +\ 1,$$

where *N* is the sample size of the GWAS, $l(j,c)$ is the LD score of SNP *j* with respect to annotation *c*, and r_{jk}^2 is the true, e.g. population-wide, genetic correlation of SNPs *j* and *k*. We define enrichment of an annotation as the proportion of heritability explained by the annotation divided by the average value of the annotation across the *M* common (MAF $\geq$ 0.05) SNPs. Enrichment may be computed for binary or probabilistic annotations according to the equation below, where $h_g^2(c)$ is the h2 explained by SNPs in annotation *c*:

$$Enrichment = \frac{h_g^2(c)\,/\,h_g^2}{\sum_j a_c\,(j)\,/\,M} = \frac{\sum_j a_c\,(j)\hat{\tau}_c\,/\,\sum_j\sum_c a_c\,\,(j)\hat{\tau}_c}{\sum_j a_c\,(j)\,/\,M}.$$

Since τ_c is not comparable between annotations or traits, τ_c^* is defined as the per-annotation standardized effect size, or the proportionate change in per-SNP h2 associated with a one standard deviation increase in the value of the annotation[34]. τ_c^* is a function of the standard deviation of the annotation *c*, *sd(c)*, the trait-specific SNP-heritability estimated by LDSC, h_g^2, and the total number of reference common SNPs used to compute h_g^2, *M* = 5,961,159 in Europeans (EUR) and 5,469,053 in East Asians (EAS):

$$\tau_c^* = \frac{sd(c)\tau_c}{h_g^2/M}.$$

τ * captures the unique contribution of an annotation to capturing h2 in the S-LDSC model, conditional on other provided annotations. Specifically, a τ * of 0, means that the annotation does not change per-SNP h2, a strongly negative τ * means that membership to the categorical annotation decreases per-SNP h2, and a strongly positive τ * means that membership to the annotation increases per-SNP h2. The significance of τ * is computed based on a test of how different from 0 the τ * is. We emphasize that enrichment does not quantify effects that are unique to a given annotation, whereas τ * does. When conditionally comparing two annotations, say A and B, in a joint S-LDSC model, both annotations may have similar enrichments if they are highly correlated. However, the τ * for the annotation with greater true causal variant membership will be larger and more significantly positive. Previous work has reported that the threshold for impactful values of $|\tau *|$ is approximately 0.24[34].

Each S-LDSC analysis conditions IMPACT annotations on 69 baseline annotations, a subset of the 75 annotations referred to as the baseline-LD model[34]; we removed 6 annotations including T cell enhancers, since IMPACT T cell-state annotations are likely correlated. The 69 annotations consist of 53 cell-type-nonspecific annotations[24], which include histone marks and open chromatin, 10 MAF bins, and 6 LD-related annotations[34] to assess if functional enrichment is cell-type-specific and to control for the effect of MAF and LD architecture. Consistent inclusion of MAF and LD associated annotations in the baseline model is the standard recommended practice of using S-LDSC.

Fine-mapped RA posterior probability enrichment in IMPACT regions. For each of 20 chosen RA-associated loci[56], we computed the enrichment of posterior probabilities in the top 1% of cell-state-specific IMPACT regulatory elements. For each RA-associated locus *l,* we define

$$Enrichment = \frac{\sum_{i}^{M_l} P_c(i)}{\sum_{i}^{M_l} 1/M_l} ,$$

where $P_c(i)$ is the posterior causal probability of SNP *i*, such that *i* belongs to the top 1% of the cell-state-specific IMPACT annotation *c*, M_l is the number of SNPs in locus *l* for which we previously computed a posterior probability[56]. The denominator represents the null hypothesis that each SNP in a locus is equally causal. We computed the average of these enrichment values over the 20 RA-associated loci. We assessed significance based on comparison to 10,000 permutation distributions, designed by computing an average enrichment value over these 20 loci, in which random posterior probabilities (of the same quantity M_l) were selected.

Results

IMPACT accurately predicts transcription factor binding

The IMPACT model assumes that cell-state-specific TF binding sites and related regulation may be characterized by a quantitative epigenomic signature. If this is true, IMPACT might predict cell-state-specific genome-wide TF occupancy with high accuracy, which has proven to be a challenging task (see ENCODE-DREAM challenge in **Web Resources**), leading to a diverse set of TF binding prediction strategies[35,36,62,63]. To test this model assumption, we used IMPACT to predict regulatory elements based on experimental binding identified via ChIP-seq of eight tested TFs assayed in eight different cell-states: T-BET, GATA3, STAT3, FOXP3, REST, HNF4A, TCF7L2, and RNA polymerase (Pol) II in CD4+ Th(T helper)1, CD4+ Th2, CD4+ Th17, CD4+ Treg (T regulatory), fetal brain, liver, pancreatic, and lymphocytic cells, respectively[32,37–41,46] (see **Material and Methods**). We observe that IMPACT predicts TF occupancy with high accuracy across 8 tested TFs. The average area under the precision-recall curve (AUPRC) over 50 random sampling trials is 0.81 (s.e. 0.07), computed via 10-fold cross validation on 80% of data, with AUPRC evaluated on the withheld 20%, **Figure 2-2A**. We additionally evaluate IMPACT using Matthew's correlation coefficient (MCC), mean MCC 0.70 (s.e. 0.08), and show full precision-recall curves (**Figure A-5**). Next, we compared IMPACT TF binding prediction performance to several recent state-of-the-art methods MocapG[35], MocapS[35], and Virtual ChIP-seq[36]. Briefly, MocapG is an unsupervised TF binding prediction method that models "cut counts" from cell-type-specific open chromatin (DNase-seq) with negative binomial distributions. MocapS is a supervised sparse logistic regression approach that predicts TF binding using cell-type-specific DNase-seq cut count modeling from MocapG, TF footprint scores from the same DNase-seq data, conservation scores, GC content, CpG island information, sequence mappability scores, and distance to nearest TSS. Virtual ChIP-seq is a multi-layer perceptron that predicts TF binding, which similarly uses conservation, cell-type-specific DNase-seq, but also leverages cell-type-specific gene expression from RNA-seq and TF-specific ChIP-seq data over a range of cell types and cell lines. While benchmarking, each method had access to the same training and testing data to ensure fair comparison. We observe that on average, across the 8 tested TFs, IMPACT outperforms all 3 methods: AUPRC IMPACT > MocapG (all $P <$ 1.5e-16, student's t-test; 0.23 average increase in AUPRC), IMPACT > MocapS (all $P <$ 5.4e-30, except for FOXP3 ($P =$ 0.15); 0.24 average increase in AUPRC), IMPACT > Virtual ChIP-seq (all $P <$ 8.5e-98; 0.62 average increase in AUPRC) (**Figure 2-2A**). We note that using Virtual ChIP-seq we were only able to predict binding for GATA3, REST, and Pol II due to data limitations. In light of this, we predicted Pol II binding in 6 additional cell types: sigmoid colon, fibroblast, left ventricle heart, liver, pancreas, and stomach. We observed that on average, IMPACT outperforms Virtual ChIP-seq according to the AUPRC (all $P <$ 4.9e-38, student's t-test; 0.48 average increase in AUPRC) (**Figure 2-2B**). We additionally used MCC as a metric to compare TF binding prediction performance, in which IMPACT also on average outperforms the competing methods (all $P <$ 2.0e-39 for MocapG, 0.30 average increase in MCC; all $P <$ 1.2e-22 for MocapS, 0.29 average increase in MCC; all $P <$ 1.2e-77 for Virtual ChIP-seq, 0.49 average increase in MCC), with the following exceptions: MCC FOXP3 IMPACT < MocapG ($P <$ 4.3e-18), MocapS ($P <$ 3.9e-19) (**Figure A-6**).

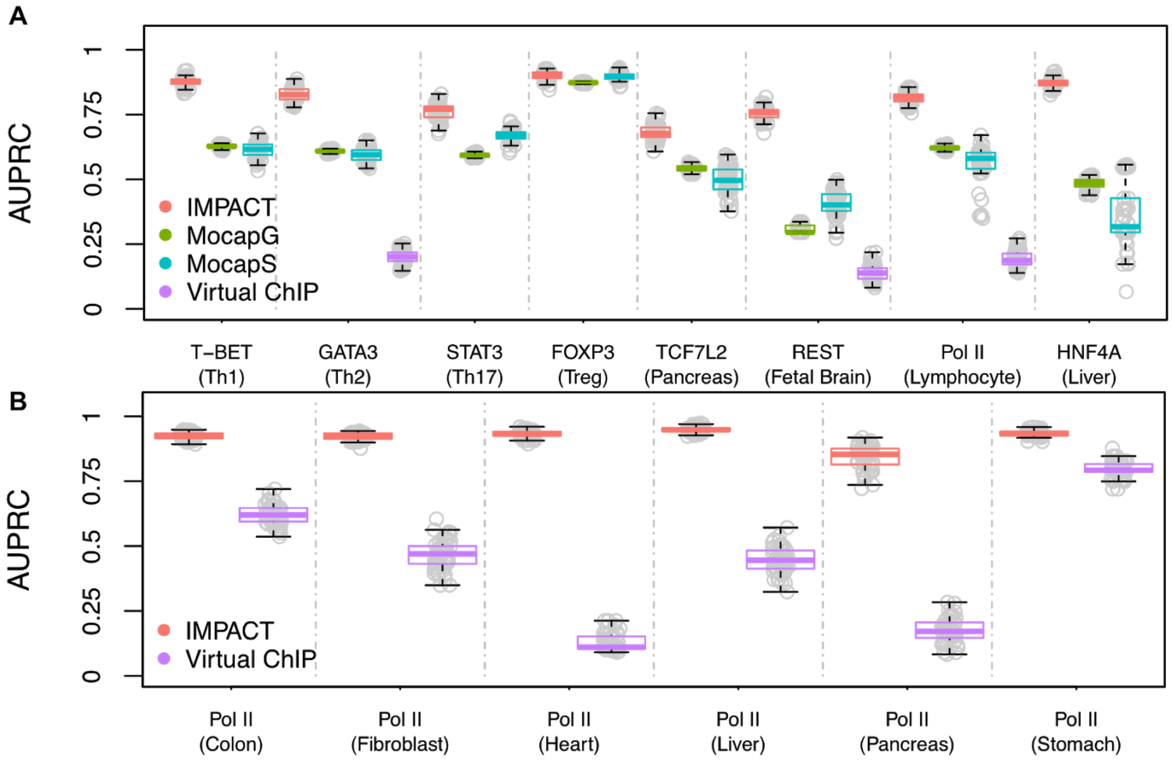


Figure 2-2. IMPACT outperforms state-of-the-art TF binding prediction methods. (a) IMPACT outperforms MocapG, MocapS, and Virtual ChIP-seq in predicting cell-state-specific TF binding across 8 TFs, illustrated by AUPRCs on the same training and testing data across 50 trials, with the exception of the MocapS model for FOXP3. (b) Prediction of Pol II binding in 6 cell types reveals that IMPACT outperforms Virtual ChIP-seq.

Genome-wide IMPACT regulatory annotations

For each of the 8 tested TFs, we created genome-wide IMPACT regulatory annotations. Focusing on the four CD4+ T cell-state IMPACT annotations, we illustrate that IMPACT regulatory element probabilities vary dynamically within TF ChIP-seq peaks near canonical CD4+ T cell-state genes. This reflects the high resolution information that is gained by integrating hundreds of epigenomic and sequence annotations (**Figure 2-3A**, **Figure A-7**). Furthermore, we observe that the most heavily weighted features from the logistic regression, indicating TF binding, include cell-state-specific open chromatin and activating histone modifications, as expected (**Figure 2-3B**). When training on entire ChIP-seq peaks rather than motif sites within peaks, top weighted features generally have less relevant cell-state-specificity (**Figure A-8**), possibly due to high correlation of ChIP-seq signal between CD4+ T cell-states. For example, most regulatory elements across CD4+ T cell-states may be near similar target genes. While the motif site regions used to train each model are TF-specific, independent and non-overlapping, we still observe relatively high correlations between CD4+ T cell-state IMPACT annotations compared to the epigenomic annotations used to train the models (**Figure A-9**).

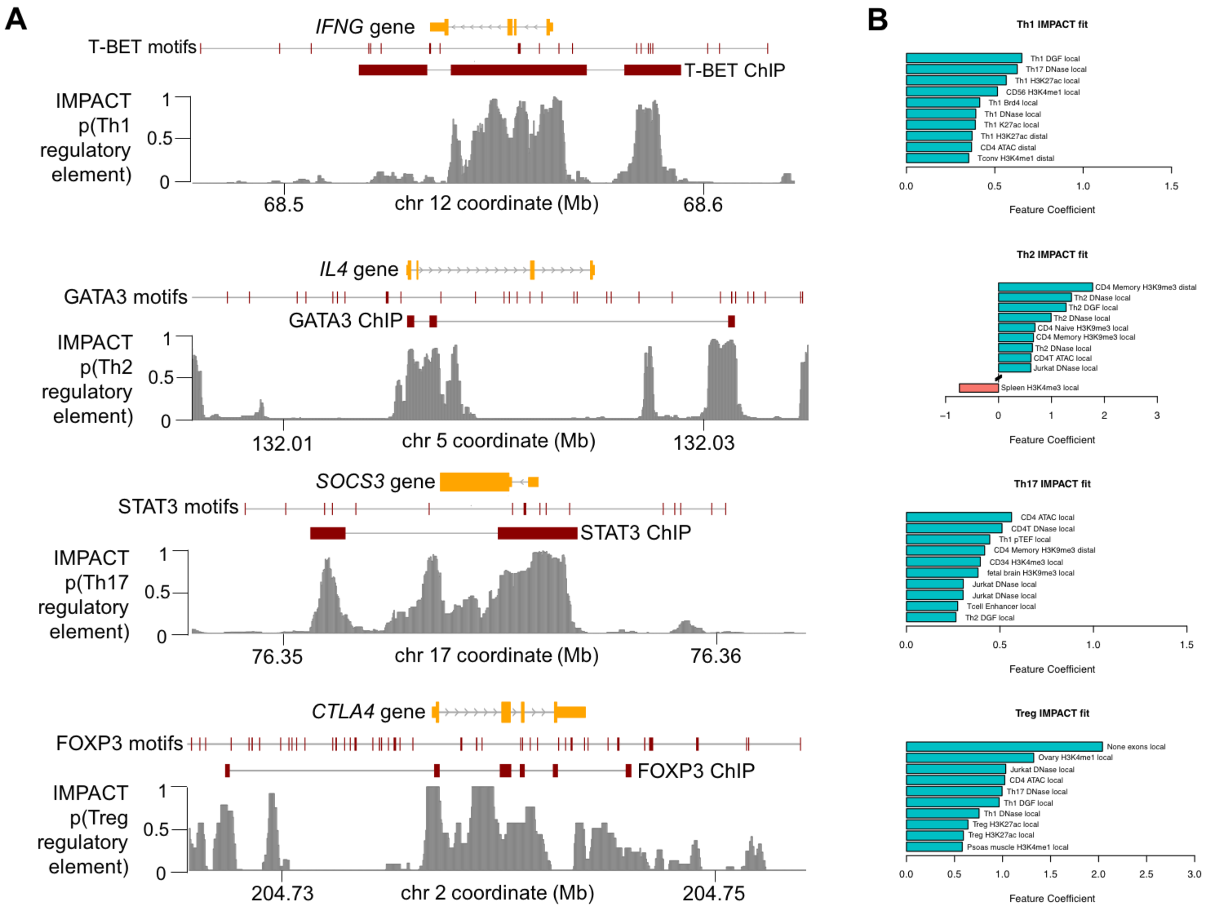


Figure 2-3. IMPACT genome-wide regulatory tracks. (a) Cell-state-specific regulatory element IMPACT predictions for canonical target genes of T-BET, GATA3, STAT3, and FOXP3. (b) Highly weighted features of Th1, Th2, Th17, and Treg IMPACT annotations.

The IMPACT epigenomic feature library contains 515 features across many cell and assay types but the importance of annotation categories is not immediately clear. To this end, we systematically removed categories of annotations and retrained TF/cell-state models (**Figure A-10**). First, we observed that TF binding predictive performance significantly decreases upon removal of cell-type-specific features for four of seven TFs (all *P* < 8.1e-4, student's t-test). For the three TFs with no significant decrease in performance, this result suggests that presence of annotations from biologically similar cell-states may be sufficient to train a high-performing IMPACT model, without requiring annotations specifically assayed in the target cell-state. Second, using just histone modification tracks resulted in significantly decreased performance on average (all *P* < 6.3e-06), while using just open chromatin tracks led to decreased performance for 5 of 7 tested TFs (all *P* < 2.1e-05) and did not significantly affect the performance for STAT3 and FOXP3. Third, we observed significantly lower performance when restricting to cell-type-specific H3K4me1 (all *P* < 2.0e-13), except for STAT3 where we observe significantly higher performance (*P* < 2.5e-44), suggesting that using cell-type-specific features only are generally less informative than a diversity of cell types and assay types. Fourth, we observed that using only cell-type-specific open chromatin results in significantly lower performance for T-BET, TCF7L2, and HNF4A (all *P* < 4.9e-17), while, for GATA3 and REST, performance improved (both *P* < 4.5e-3); no comparison could be made for STAT3 or FOXP3 because there were no Th17 or Treg open chromatin annotations to begin with. From this, we learn that integration of diverse cell types and assays generally leads to improved predictive performance. In the case of GATA3, STAT3, and REST, where the use of only cell-type-specific annotations resulted in improved performance over the canonical IMPACT model, such models may overfit to training data and misrepresent true TF binding patterns genome-wide. Therefore, further assessment is necessary, specifically involving training and testing across multiple datasets from the same cell type, which was not possible in this study due to scarcity of primary cell TF ChIP-seq data.

Improved enrichment of gene expression causal variation

We developed IMPACT to model regulation specific to a functional cell-state, the most general of which may be active cellular transcription. Expression quantitative trait loci (eQTLs) are genetic variations that modulate transcription[64]. Most cis eQTLs map to TSS and promoter annotations, and more rarely to the 5' UTR[65]. We hypothesized that an IMPACT annotation tracking active transcription, trained on RNA polymerase (Pol) II binding sites, would capture cis eQTL causal variation better than the most strongly enriched canonical eQTL-related annotations.

We obtained SNP-level summary statistics from three independent sources: first, from a large and previously published eQTL analysis on 3,754 peripheral blood samples[49]; second, from GTEx V7 across 6 tissue types (average sample size = 231): transformed fibroblasts, stomach, liver, left ventricle heart, sigmoid colon, and pancreas; and third, from a CD4+ T cell eQTL analysis on 103 East Asian individuals[50]. We then used IMPACT to annotate SNPs tested in the eQTL analysis with RNA Pol II specific regulatory element probabilities, separately for each tissue or cell type. In this analysis, we were limited by the availability of Pol II ChIP-seq, for which there is an abundance of tissue-specific data but rarely more specific cell-type level data.

While tissues may contain many different cell types, we expect IMPACT to learn an epigenomic signature as general or as specific as the training data provided. For the peripheral blood IMPACT annotation, we combined sites of Pol II binding in both T cells and B cells, the predominant cell populations of peripheral blood, and trained a single IMPACT model. Next, we computed a genome-wide enrichment (see Material and Methods) of chi-squared cis eQTL association statistics, averaged over all genes with at least one significant eQTL, across Pol II IMPACT, Pol II ChIP-seq, and several sequenced-based annotations, such as TSS windows, promoters, and enhancers. We observed that on average Pol II IMPACT across all cell types was more enriched for chi-squared association than the Pol II ChIP-seq used for training (paired t-test *P* < 7.7e-4, 7.3% average increase in enrichment). To compute this, we thresholded each Pol II ChIP-seq dataset by peak score, considering 11 uniformly spaced cutoffs, ranging from highest scoring ChIP-seq peaks to lowest scoring, while still significant, peaks. To directly compare with IMPACT, we appropriately thresholded each Pol II IMPACT annotation by matching on size, e.g. the genome-wide proportion of SNPs annotated (**Figure A-11**). We also computed enrichment for IMPACT at 5 other annotation size thresholds (0.5%, 1%, 2.5%, 5%, and 10%), which resulted in larger enrichments than achievable by any thresholding of the ChIP-seq data. Furthermore, we observe that Pol II IMPACT captures more chi-squared association than sequenced-based functional annotations (student's t-test p<4.8e-4) with the highest performing IMPACT annotations providing a 25% average increase in enrichment over the sequenced-based annotations (**Figure 2-4**). For each of the eight tissues or cell types tested, the most enriched Pol II IMPACT annotation outperformed all sequenced-based functional annotations. Specifically, in peripheral blood, Pol II IMPACT introduced a 1.7x enrichment (permutation *P* < 1e-3), corresponding to a 24% average increase in enrichment compared to the tested sequence-based functional annotations. Similarly, for transformed fibroblasts, Pol II IMPACT introduced a 1.7x enrichment (30% increase); for stomach, a 1.5x enrichment (22% increase); for liver, a 1.5x enrichment (25% increase); for left ventricle heart, a 1.5x enrichment (22% increase); for sigmoid colon, a 1.5x enrichment (22% increase); for pancreas, a 1.5x enrichment (18% increase); and for CD4+ T cells, a 1.8x enrichment (41% increase).

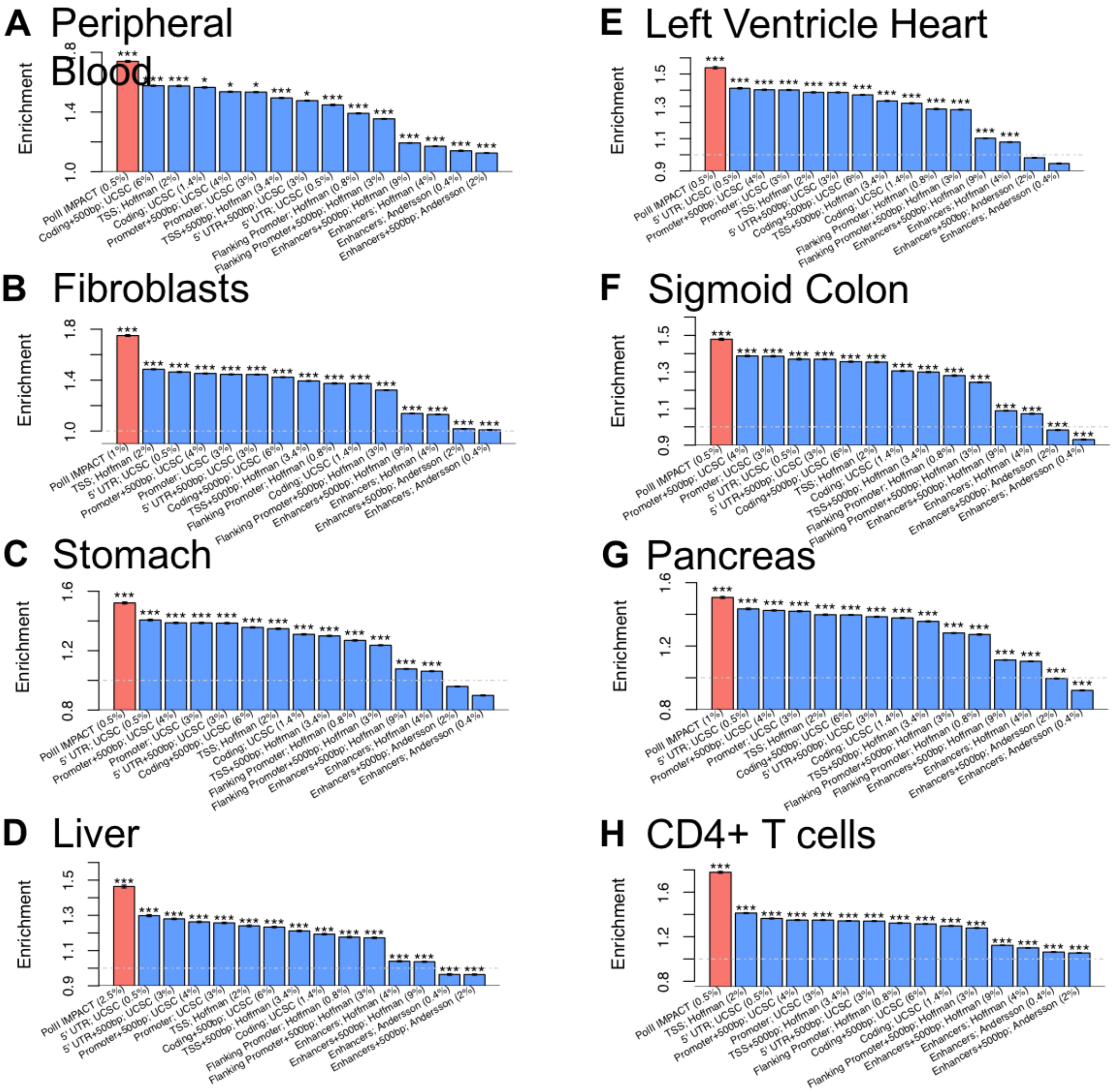


Figure 2-4. Pol II IMPACT captures cis eQTL causal variation better than sequence-based annotations across 8 cell and tissue types. Enrichment of cis eQTL chi-squared association values with Pol II IMPACT annotations, created for peripheral blood (a), fibroblasts (b), stomach (c), liver (d), left ventricle heart (e), sigmoid colon (f), pancreas (g), and CD4+ T cells (h), highlighting top performing IMPACT annotation compared to enrichments of sequence-based functional annotations. Values in parentheses after annotation name are the average annotation value across all common variants, e.g. the effective size of the annotation. * denotes permutation $P < 0.05$, ** permutation $P < 0.01$, *** permutation $P < 0.001$. Intervals at the top of each bar represent the 95% confidence interval of the enrichment estimate.

Improved capture of rheumatoid arthritis causal variation

We previously hypothesized that IMPACT annotations of pathogenic cell-states would more precisely capture polygenic trait h2, compared to regulatory annotations that don't resolve cell-states. Testing this hypothesis requires a polygenic trait with a well-studied disease-driving cell type. Genetic studies of rheumatoid arthritis (RA), an autoimmune disease that attacks synovial joint tissue leading to permanent joint damage and disability[66], have suggested a critical role by CD4+ T cells[13,14,16,17,24,25,67–69]. However, CD4+ T cells are extremely heterogeneous: naive CD4+ T cells may differentiate into memory T cells, and then into effector T cells including Th1, Th2, and Th17 and T regulatory cells, requiring the action of a limited number of key transcription factors (TFs): T-BET or STAT4, GATA3 or STAT6, STAT3 or RORt, FOXP3 or STAT5, respectively[70]. As these CD4+ T effector cell-states contribute to RA risk[14,17,24], we hypothesized that CD4+ T cell-state-specific IMPACT regulatory element annotations would better capture RA h2 than annotations that generalize CD4+ T cells and ignore the differential functionality of effector cell-states.

To this end, we built IMPACT annotations in four CD4+ T cell-states, Th1, Th2, Th17, and Treg. We then integrated S-LDSC[24] with publicly available European (EUR, N = 38,242)[24,51] and East Asian (EAS, N = 22,515)[52] RA GWAS summary statistics to partition the common SNP h2 of RA. We use two metrics to evaluate how well our IMPACT annotations capture RA h2: enrichment and per-annotation standardized effect size, τ^* (see **Material and Methods**). Briefly, enrichment is defined as the proportion of h2 divided by the genome-wide proportion of SNPs in the annotation, and τ^* is defined as the proportionate change in per-SNP h2 associated with a one standard deviation increase in the value of the annotation[34].

We observe that each CD4+ T cell-state-specific IMPACT annotation is significantly enriched with RA h2 in both EUR and EAS populations (average enrichment = 20.05, all $P < 1.9\text{e-}04$, **Figure 2-5A**, **Table A-5**). Furthermore, we find that τ^* is significantly positive for all CD4+ T IMPACT annotations separately conditioned on the cell-type-nonspecific baseline-LD annotations (all $P < 2.1\text{e-}03$, **Figure 2-5B, Figure A-12**), supporting the CD4+ T cell-specific role in RA. We then selected the top 5% of regulatory SNPs according to each CD4+ T IMPACT annotation and find that all the CD4+ T cell-state annotations explain a large proportion of RA h2, but the Treg annotation explains the greatest proportion, capturing 85.7% (s.e. 19.4%, enrichment $P < 1.6\text{e-}5$) of RA h2 meta-analyzed between both EUR and EAS populations (**Figure 2-5C**). Furthermore, we observe that the top 9.8% of CD4+ Treg IMPACT regulatory elements, consisting of all SNPs with a non-zero annotation value, capture 97.3% (s.e. 18.2%, enrichment $P < 7.6\text{e-}7$) of RA h2 in EUR. This powerful result is the most comprehensive explanation for RA h2, to our knowledge, to date.

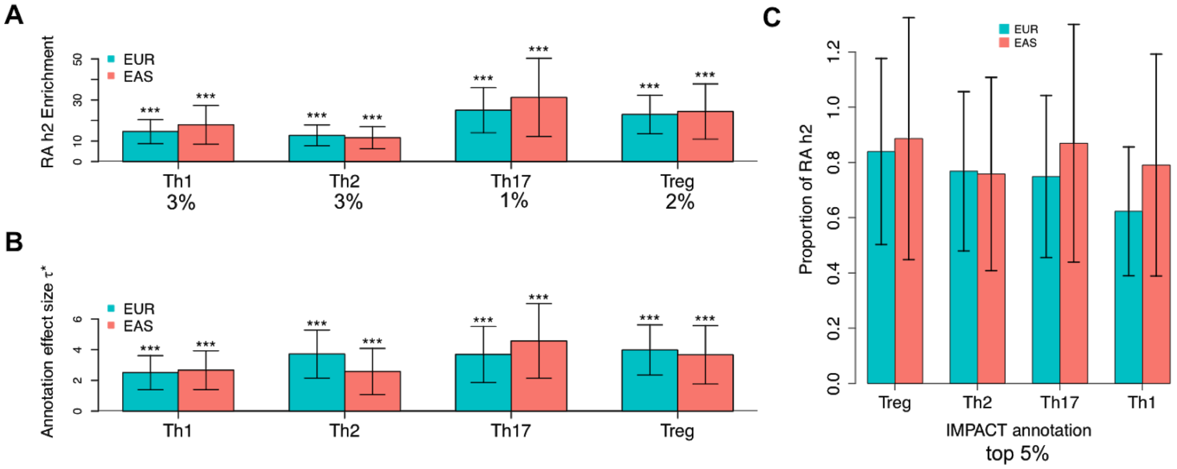


Figure 2-5. CD4+ T cell-state IMPACT annotations are strongly enriched for RA heritability. (a) Enrichment of RA h2 in CD4+ T IMPACT for EUR and EAS populations. Values below cell-states are the average annotation value across all common (MAF ≥ 0.05) SNPs, e.g. the effective size of the annotation. (b) Standardized annotation effect size (τ *) of each annotation separately conditioned on annotations from the baseline-LD model. For panels a and b, *** denotes $P < 0.001$. (c) Proportion of total causal RA h2 explained by the top 5% of SNPs in each IMPACT annotation. For all panels, 95% CI represented by black lines.

We then assessed if CD4+ T IMPACT annotations offered improved enrichments of RA h2 compared to canonical CD4+ T cell functional annotations, using S-LDSC and EUR RA summary statistics (**Figure 2-6A, Figure A-13**). Here, we highlight our comparison of the CD4+ Treg IMPACT annotation to FOXP3 binding motif sites, genome-wide FOXP3 ChIP-seq, the “Averaged Tracks” annotation, which assigns each SNP a value proportional to the number of overlapping IMPACT epigenomic features, the five largest τ * CD4+ T cell-specific histone mark annotations[24], the five largest τ * CD4+ T cell-specifically expressed gene sets[25], and CD4+ T cell super enhancers[71]. We observe that the CD4+ Treg IMPACT annotation (enrichment = 22.9, s.e. 4.8, $P < 5.2e-08$) is significantly more enriched ($P < 0.05$) for RA h2 than the FOXP3 motif site annotation (enrichment not significantly different from 0), the “Averaged Tracks” annotation (enrichment = 7.0, s.e. 1.4), all CD4+ T cell-specifically expressed gene sets (average enrichment = 2.9, s.e. 0.8), and CD4+ T cell super enhancers (enrichment = 8.1, s.e. 1.3). On the other hand, the FOXP3 ChIP-seq annotation (enrichment = 173.3, s.e. 58.3), which is used to train the CD4+ Treg IMPACT model, is more strongly enriched ($P < 0.05$) for RA h2 than the CD4+ Treg IMPACT annotation itself. We additionally created functional annotations representing the overlap of TF ChIP-seq with TF motif sites, as such a combination might improve the enrichment observed for TF ChIP-seq alone. However, these annotations are very small (average annotation size = 0.004% of SNPs) and resulted in non-significant enrichments in the S-LDSC framework. Finally, we observe that all compared CD4+ T cell histone mark annotations are similarly enriched for RA h2, relative to the CD4+ Treg IMPACT annotation (23.4x on average compared to 22.9x, respectively). We note that the average RA h2 captured by these CD4+ T histone mark annotations, ranging in size from 1-3% of SNPs, is 42.3%; and, the average RA h2 captured by these CD4+ T specifically expressed gene set annotations, ranging in size from 11-13% of SNPs, is 36.4%. In terms of total RA h2 explained by a single annotation, these values pale in comparison to the 85.7% of RA h2 captured by the top 5% of SNPs in the Treg IMPACT annotation.

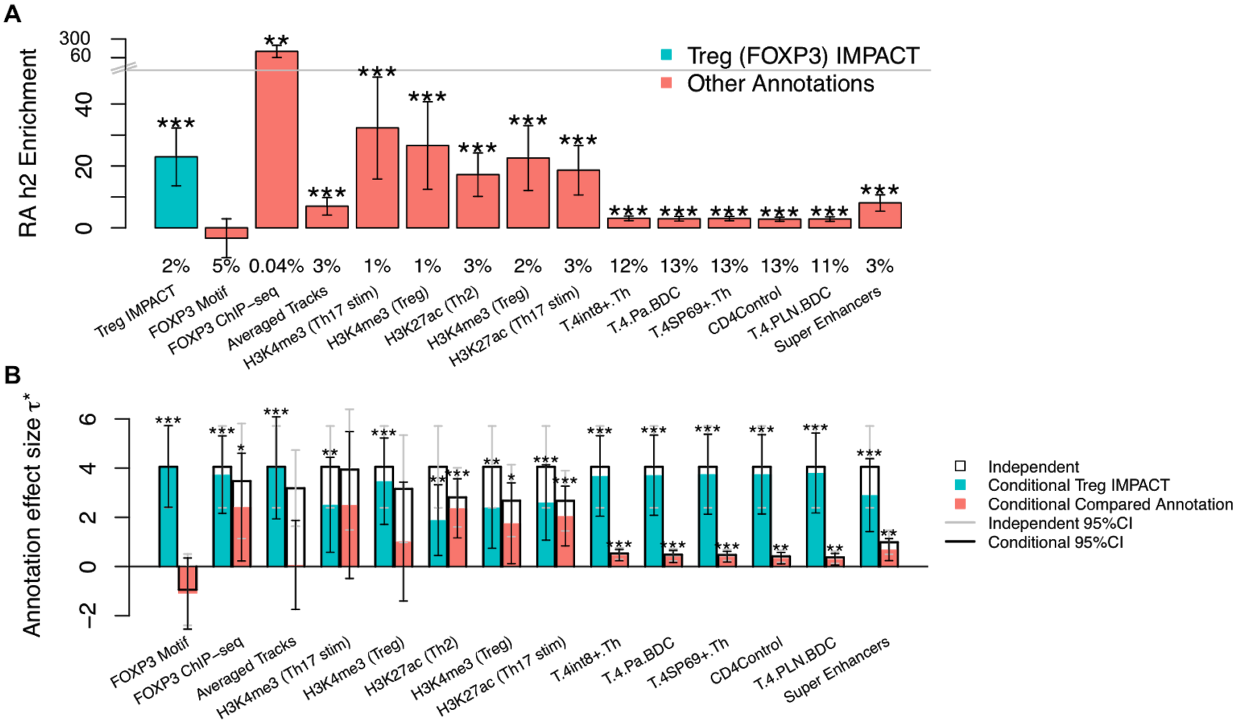


Figure 2-6. CD4+ Treg IMPACT annotation significantly captures RA heritability conditional on strongly enriched CD4+ T cell regulatory annotations. (a) RA h2 enrichment of the CD4+ Treg IMPACT annotation and compared T cell functional annotations. Values below cell-states represent the effective size of the annotation. From left to right, we compare Treg IMPACT to genome-wide FOXP3 motif sites, FOXP3 ChIP-seq, the "Averaged Tracks" annotation, which assigns each SNP a value proportional to the number of overlapping IMPACT epigenomic features, the top 5 cell-type-specific histone modification annotations, in terms of independent τ *, the top 5 cell-type-specifically expressed gene sets, in terms of independent τ *, and T cell super enhancers. (b) CD4+ Treg IMPACT annotation standardized effect size (τ *, teal) conditional on other T cell related functional annotations (coral). τ * for independent analyses are denoted by the top of each black bar, as a reference for the conditional analyses, denoted by the top of each colored bar. For panels a) and b), * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Next, in order to quantify annotation-specific effects of capturing RA h2, we computed the per-annotation standardized effect size, τ *, of each annotation from the previous analysis, conditioned on baseline-LD annotations. We then separately conditioned each CD4+ T cell-state IMPACT annotation jointly on the compared annotations and baseline-LD annotations. Larger and more significantly positive τ * identifies the annotation that better captures RA h2. We observe that the τ * of both CD4+ Treg and Th2 IMPACT annotations are larger and more significantly positive (all Treg τ * > 1.9, *P* < 5.0e-3; all Th2 τ * > 1.7, *P* < 0.01) than compared T cell annotations, excluding H3K27ac in Th2 cells, illustrated by taller teal bars than coral bars (**Figure 2-6B, Figure A-13**). Here, we specifically highlight the CD4+ Treg IMPACT annotation; although the FOXP3 ChIP-seq annotation was more strongly enriched for RA h2 than CD4+ Treg IMPACT, the τ * of the IMPACT annotation is larger and more significantly positive. Overall, these results suggest that IMPACT annotates areas of concentrated RA h2 that other T cell regulatory annotations do not.

IMPACT annotation effect sizes across 42 polygenic traits

We next applied our CD4+ T IMPACT annotations to 41 additional polygenic traits[34,53,54] and observed consistently significantly positive per-annotation standardized effect sizes, τ *, for immune-mediated traits, such as Crohn's, "all autoimmune disease", respiratory ear/nose/throat, and "allergy and eczema" (mean τ * = 3.2; all *P* < 5.9e-4, *P* < 1.9e-5, *P* < 3.6e-3, *P* < 1.7e-3, respectively), and for several blood traits, eosinophil and white blood cell counts (mean τ * = 2.5; all *P* < 1.6e-11, *P* < 0.02, respectively), but not for non-immune-mediated traits (**Figure 2-7A, Table A-6**). We then created several different cell-state-specific IMPACT annotations targeting h2 in a range of traits; and, we highlight a few examples. For a liver IMPACT annotation, trained on HNF4A[32] (hepatocyte nuclear factor 4A), τ * is positive for liver-associated traits[46,72] LDL and HDL (mean τ * = 2.0; $P < 0.02$, $P < 1.2e\text{-}3$, respectively). For a macrophage IMPACT annotation, trained on IRF5[42], τ * is positive for some immune-mediated and blood traits (mean τ * = 2.8, all $P < 8.2e\text{-}3$) and intriguingly also for schizophrenia (τ * = 0.9, $P < 4.9e\text{-}5$), supported by studies implicating a putative MHC association[73]. Finally, for a CD4+ Treg IMPACT annotation, trained on STAT5[41], an alternative key TF for Tregs, the values of τ * across all traits resemble that of FOXP3. This suggests that IMPACT is capturing RA polygenic h2 by annotating loci important to Treg function, rather than TF-specific loci. To ensure that IMPACT annotations were an improvement over the original ChIP-seq used to train each model, we compute τ * across the same 42 traits for annotations created from the training TF ChIP-seq data (**Figure 2-7B**). We observe fewer significant effect sizes, with the exception of stronger τ * in the T-BET ChIP-seq compared to the T-BET (Th1) IMPACT annotation, first identified in the conditional analysis in **Figure A-13**. Overall, this suggests that IMPACT is a promising strategy to identify complex trait associated regulatory elements across a range of cell-states.

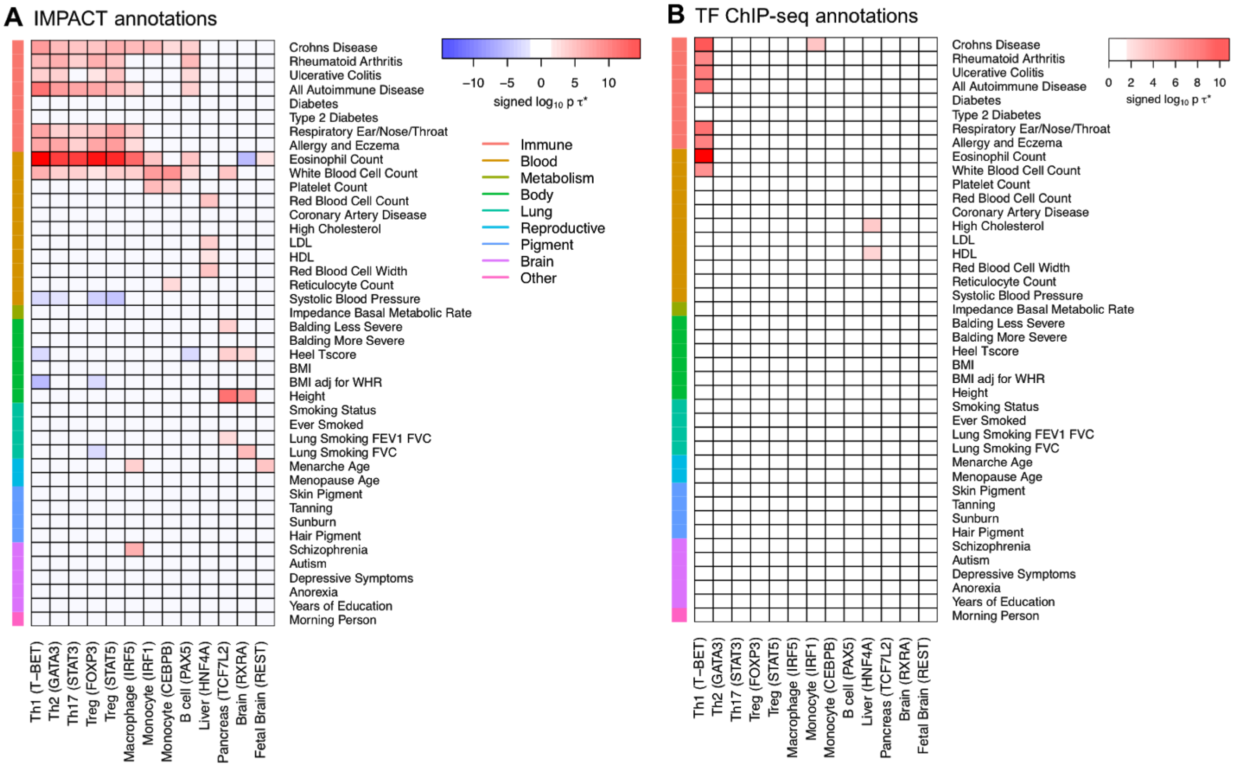


Figure 2-7. IMPACT cell-state-specific regulatory element annotation effect sizes across 42 polygenic traits. (a) Signed $\log_{10}$ *P* values of τ^* for 42 traits across 13 cell-state-specific IMPACT annotations, capturing h2 in distinct sets of complex traits, shown by significantly positive τ^*. Each IMPACT annotation is described by its target cell-state and key TF used for training in parentheses. (b) Signed $\log_{10}$ *P* values of τ^* for 42 traits across annotations representing the TF ChIP-seq used to train the corresponding IMPACT annotations. ChIP-seq annotations are described by the cell-state in which the particular TF (in parentheses) was assayed. For both panels, color shown only if *P* value of $\tau^* < 0.025$ after multiple hypothesis correction.

A priori functional characterization of variants

We next hypothesized that improved genomic annotation provided by IMPACT might inform functional variant fine-mapping. Using a GWAS of 11,475 European RA cases and 15,870 controls[57], an independent study from the European RA summary statistics used in our h2 analyses, our group recently fine-mapped a subset of 20 RA risk loci, each with a manageable number of putatively causal variants, and created 90% credible sets of these SNPs[56]. We computed the enrichment of fine-mapped causal probabilities across these 20 loci in the top 1% of our CD4+ T cell-state-specific IMPACT annotations (see Material and Methods). We found that the Treg annotation is significantly enriched (2.87, permutation $P < 1.8e-02$) while other annotations are not (**Table A-7**). The Treg IMPACT annotation may thus be useful to prune putatively causal RA variants. Furthermore, we observe uniquely high Treg enrichment in the *BACH2* and *IRF5* loci (16.2 and 8.1, respectively, **Figure 2-8A**), suggesting putatively causal SNPs in these loci may function in a Treg-specific context.

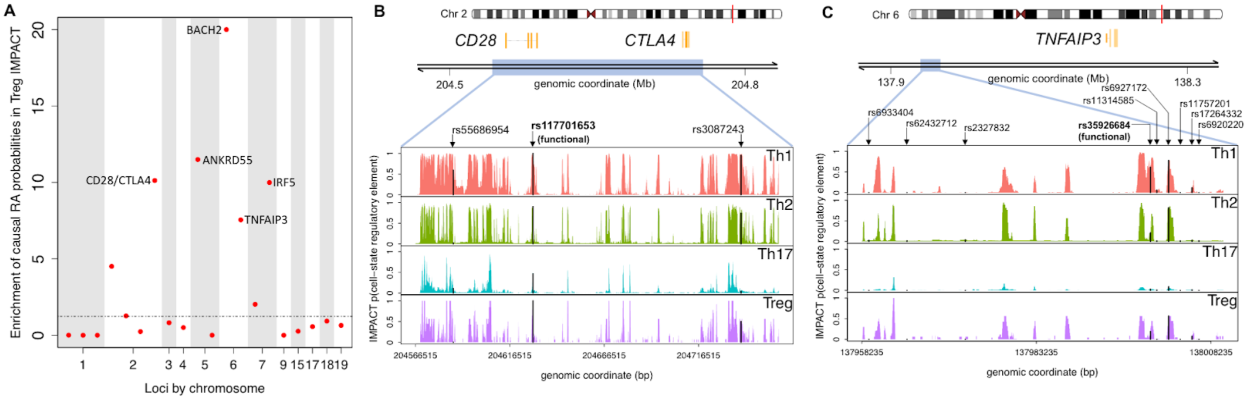


Figure 2-8. IMPACT a priori identifies variants with measured functionality. (a) Enrichment of posterior probabilities of putatively causal RA SNPs in the top 1% of SNPs with CD4+ Treg regulatory element probabilities highlights the *BACH2*, *ANKRD55*, *CTLA4/CD28*, *IRF5*, and *TNFAIP3* loci. (b,c) IMPACT regulatory element probabilities (black) at putatively causal SNPs with experimentally validated differential enhancer activity (bolded) and other 90% credible set SNPs (unbolded) at two RA-associated loci, *CTLA4/CD28* and *TNFAIP3*.

In the same study, our group observed both differential binding of CD4+ T nuclear extract via EMSA and differential enhancer activity via luciferase assays at two credible set SNPs, narrowing down the list of putatively causal variants in the *CD28/CTLA4* and *TNFAIP3* loci[56]. We observed that both variants with functional activity were located at high probability IMPACT regulatory elements, suggesting that IMPACT may be used to narrow down credible sets to reduce the amount of experimental follow up. First, at the *CD28/CTLA4* locus, IMPACT predicts high probability regulatory elements across the four CD4+ T cell-states at the functional SNP rs117701653 and lower probability regulatory elements at other credible set SNPs rs55686954 and rs3087243 (**Figure 2-8B**). Second, at the *TNFAIP3* locus, we observe high probability regulatory elements at the functional SNP rs35926684 and other credible set SNP rs6927172 (**Figure 2-8C**) and do not predict regulatory elements at the other 7 credible set SNPs. The CD4+ Th1 specific regulatory element at rs35926684 suggests that this SNP may alter gene regulation specifically in Th1 cells and hence, we suggest any functional follow-up be done in this cell-state. Fewer than 11% of the credible set SNPs in the other 18 fine-mapped loci have high IMPACT cell-state-specific regulatory element probabilities (**Figures A-14 to A-16**).

Discussion

In summary, we assume that cell-state-specific regulation may be characterized by an epigenomic signature that may be captured by the cell-state-specific binding sites of a single key TF. To this end, we designed IMPACT to predict cell-state-specific regulatory elements based on epigenomic and sequence profiles of experimental cell-state-specific TF binding by performing a logistic regression on 515 such features. We specifically chose not to employ a deep learning approach in order to retain interpretability of learned annotation weights. Knowledge of which epigenomic or sequence feature annotations are most informative for predicting transcriptional regulation, which varies among cell-states, can guide where experimental assay resources might be invested to learn more about the regulome.

We demonstrated the versatility of IMPACT as a genome annotation strategy with several compelling applications. First, we observed that the robust epigenomic footprint of TF binding sites allows for accurate binding prediction. Furthermore, IMPACT outperformed three state-of-the-art methods, MocapG, MocapS, and Virtual ChIP-seq which use a compendium of sequence-based, open chromatin and gene expression annotations to predict cell-state-specific TF binding. We believe that this increased predictive power comes from the way in which IMPACT learns which genomic annotations are correlated with TF binding, without knowledge of the cell type or cell-state of interest. This is contrary to the compared methods where cell-type-specific DNase-seq or ATAC-seq must be provided as a reference. Moreover, IMPACT provides epigenomic annotations from a wide variety of cell types and assay types which provide complimentary information. We note that we restrict binding prediction to motif sites for each TF in a given cell type. Moreover, validation in a completely independent ChIP-seq dataset was not possible due to the scarcity of primary cell TF ChIP-seq data.

Second, using Pol II IMPACT annotations, for eight tested tissue and cell types, we more precisely captured causal variation of gene expression than by using Pol II ChIP-seq and sequence-based annotations. Our results argue that Pol II IMPACT regions better localize active promoter and proximal regulatory regions driving eQTLs than the compared canonical genomic annotations, which may be less specific due to their larger sizes and restrictive binary characterization. This suggests that IMPACT may be more effective at prioritizing causal SNP variation when fine-mapping eQTLs. These results also argue that the biological basis of eQTLs are related to Pol II binding regions, which is a refinement over previous observations that eQTL causal variation is concentrated near and around TSS and promoter regions.

Third, we more precisely captured causal variation of complex traits. Our CD4+ T IMPACT annotations capture more RA h2 than most canonical CD4+ T cell regulatory annotations. Our findings further reinforce that IMPACT annotations, as an aggregation of hundreds of regulatory annotations, are more informative than single annotations. This is exemplified by the finding that FOXP3 ChIP-seq is strongly enriched for RA h2; and, while this annotation was used as training data for IMPACT, the CD4+ Treg IMPACT annotation captured more RA h2, evident by a larger, more significant annotation effect size, τ *, in the joint analysis. Furthermore, we showed that CD4+ T cell IMPACT annotations explain similar proportions of RA heritability in both European and East Asian populations, suggesting that biological mechanisms driving RA may have similar genetic and regulatory bases in these two populations. We also demonstrated that our approach is generalizable to other trait-driving cell types by showing significantly positive τ * of IMPACT annotations for 21 of 42 tested complex traits. In particular, CD4+ T IMPACT annotations also captured significant h2 of autoimmune and immune-mediated traits, which is expected given the central role of CD4+ T cells to the immune system and perhaps shared genetic architecture of these traits. We find that h2 of intuitively brain-related traits such as schizophrenia, anorexia, and autism is not captured by brain IMPACT annotations, perhaps suggesting that more complex, cross-cell-type regulatory networks are core to the genetic risk of these traits. Rather, brain IMPACT annotations capture h2 of traits such as menarche age, smoking, and height. We note that we targeted specific polygenic traits using a priori knowledge of the cell-states that were most likely to be driving causal biology. To better refine or inform the choice of relevant cell type, we recommend integrating IMPACT with previously published approaches, such as

RolyPoly[74], which prioritizes cell types with respect to a particular trait, based on linking single cell gene expression to GWAS summary statistics. We note that S-LDSC analyses exclude the major histocompatibility complex due to its extremely high gene density and outlier LD structure, which is thought to be the strongest contributor to RA disease h2[75]. However, our work supports the notion that there is an undeniably large amount of RA h2 located outside of the MHC.

Lastly, we demonstrated that IMPACT may identify functional variants a priori and suggest the relevant cell-state contexts in which these functional variants may act. We note that disease-relevant IMPACT functional annotations may be integrated with existing functional fine mapping methods, like PAINTOR[76] or CAVIARBF[77], to assign causal posterior probabilities to variants.

We recognize several important limitations to our work. First, we have not experimentally validated the activity of any of our predicted regulatory elements. Second, predicted regulatory elements are limited to genomic regions that have been epigenetically assayed. Third, IMPACT as presented in this study, is limited to cell-states in which ChIP-seq of a key TF has been performed. Furthermore, some TFs are key regulators in more than one cell type or cell-state, which should not compromise the cell-state-specificity of the learned IMPACT annotation. We note that cell-state-specificity is not gained from the TF itself, but from the unique binding patterns of the TF in a modeled cell-state. For example, the CD4+ T cell TFs, for which we create IMPACT annotations, are also key regulators in analogous cell-states of ILCs (innate lymphoid cells)[78]. Under the assumption that these key TFs regulate different sets of genes in the analogous cell-states, cell-state-specific IMPACT annotations learned from, for example, T-BET in CD4+ Th1s should be distinguishable from an annotation learned for T-BET in ILC1s.

Due to the lack of functional data on ILCs, we were not able to test this claim. However, as more cell-state and cell type data is generated, especially on more fine resolution cellular populations, better regulatory annotations may be produced. Moreover, these new functional annotations might nominate other or more precise cellular populations, compared to the ones considered in this study, for explaining polygenic trait heritability and capturing fine-mapped causal variation. While we highlight strong enrichments of IMPACT models trained on CD4+ T cell TFs, especially FOXP3, we acknowledge that it is certainly possible that other cell types and factors play important roles that we have not explored in this study. Fourth, S-LDSC heritability analyses results may be sensitive to the size of the annotation and we recommend enforcing reasonably large annotation sizes, for example at least 0.1% of the genome (**Figure A-17**). In light of these limitations, IMPACT is an emerging strategy for identifying trait associated regulatory elements and generating hypotheses about the cell-states in which variants may be functional, motivating the need to develop therapeutics that target specific disease-driving cell-states.

Web Resources

1. S-LDSC tutorial and instructions: github.com/bulik/ldsc
2. 1000G: www.1000genomes.org
3. RA EUR summary statistics: http://plaza.umin.ac.jp/yokada/datasource/software.htm
4. RA EAS summary statistics: http://jenger.riken.jp/en/result
5. 1000G Phase 3 LD scores, CD4+ T cell specifically expressed genes (binary functional annotations): http://data.broadinstitute.org/alkesgroup/LDSCORE/
6. Immgen.tsv: https://gist.github.com/nachocab/3d9f374e0ade031c475a

7. GTEx data: https://gtexportal.org/home/
8. HOMER: http://homer.ucsd.edu/homer/motif/
9. IMPACT GitHub repository: https://github.com/immunogenomics/IMPACT

Acknowledgements

We would like to thank Eric Lander and Jesse Engreitz for helpful discussions. This work is supported in part by funding from the National Institutes of Health (NHGRI T32 HG002295, UH2AR067677, 1U01HG009088, U01 HG009379 (Price/Raychaudhuri U01), and 1R01AR063759) and the Doris Duke Charitable Foundation Grant #2013097.

Chapter 3

Improving the trans-ethnic portability of polygenic risk scores by prioritizing variants in predicted cell type regulatory elements

The material in this chapter appeared on bioRxiv on February 28, 2020 and is currently in revision at *Nature Genetics*.

Abstract

Poor trans-ethnic portability of polygenic risk score (PRS) models is an important issue caused in part by Eurocentric genetic studies and in part by limited knowledge of causal variants shared among populations. Hence, leveraging noncoding regulatory annotations that capture genetic variation across populations has the potential to enhance the trans-ethnic portability of PRS. To this end, we constructed a unique resource of 707 cell-type-specific IMPACT regulatory annotations by aggregating 5,345 public epigenetic datasets to predict binding patterns of 142 cell-type-regulating transcription factors across 245 cell types. With this resource, we partitioned the common SNP heritability of diverse polygenic traits and diseases from 111 GWAS summary statistics of European (EUR, average N=180K) and East Asian (EAS, average N=157K) origin. For 95 traits, we were able to identify a single IMPACT annotation most strongly enriched for trait heritability. Across traits, these annotations captured an average of 43.3% of heritability (sem = 2.8%) with the top 5% of SNPs. Strikingly, we observed highly concordant polygenic trait regulation between populations: the same regulatory annotations captured statistically indistinguishable SNP heritability (fitted slope = 0.98, sem = 0.04). Since IMPACT annotations capture both large and consistent proportions of heritability across populations, prioritizing variants in IMPACT regulatory elements may improve the trans-ethnic portability of PRS. Indeed, we observed that EUR PRS models more accurately predicted 21 tested phenotypes of EAS individuals when variants were prioritized by key IMPACT tracks (49.9% mean relative increase in R^2). Notably, the improvement afforded by IMPACT was greater in the trans-ethnic EUR-to-EAS PRS application than in the EAS-to-EAS application (47.3% vs 20.9%, one-tailed paired wilcoxon *P* < 0.012). Overall, our study identifies a crucial role for functional annotations such as IMPACT to improve the trans-ethnic portability of genetic data.

Introduction

An important challenge for complex trait genetics is that there is no clear framework to transfer population-specific genetic data, such as GWAS results, to individuals of other ancestries[79–81]. The importance of this challenge is accentuated by the fact that approximately 80% of all genetic studies have been performed with individuals of European ancestry, accounting for a minority of the world's population[82]. This is exacerbated by the fact that population-specific linkage disequilibrium (LD) between variants confounds inferences about causal cell types and variants (**Figure 3-1A**)[23,24,83]. GWAS have the potential to revolutionize the clinical application and utility of genetic data to the individual, exemplified by current polygenic risk score (PRS) models[18,19,83–90]. However, while the utility of PRS models relies on accurate estimation of allelic effect sizes from GWAS and benefits from genetic similarity between the target cohort and the training GWAS cohort, recent studies have explicitly observed a lack of trans-ethnic portability[18,80,81,83,91,92]. The Eurocentric GWAS bias has led PRS to be more predictive in European populations, as the largest training data comes from European GWAS[81,83,86,93,94]. As a result, variants used in European PRS tend to be more common among Europeans and less common among non-Europeans. Common variants carry greater disease predictive power which directly contributes to Eurocentric bias in PRS accuracy[81]. The trans-ethnic portability of PRS would not be as critical an issue if large GWAS were performed in all non-EUR populations. Previous studies have extensively shown that functional annotations can improve PRS models when learned and applied to the same population[95,96], by introducing biologically-relevant priors on causal effect sizes and compensating for inflation of association statistics by LD. However, the potential for functional annotations to improve trans-ethnic PRS frameworks, where the influences of population-specific LD are more profound, has not yet been extensively investigated.

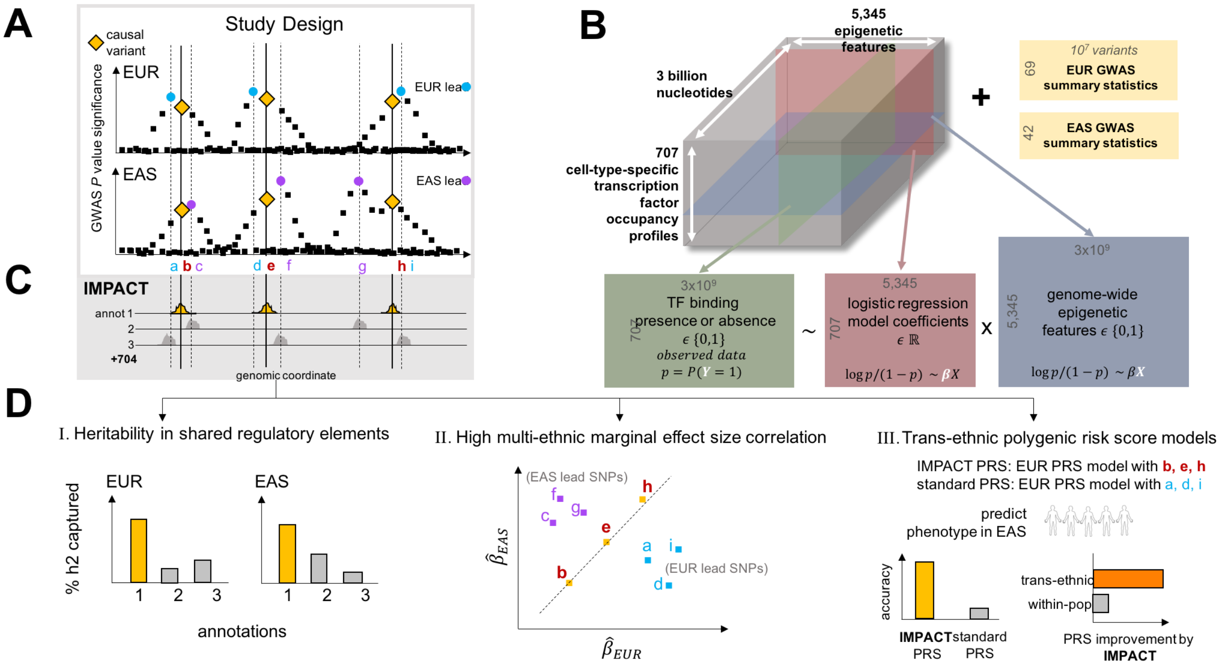

A
Study Design
causal variant
EUR
EUR lead
EAS
EAS lead
GWAS P value significance
a b c
d e f
g
h i
C
IMPACT
annot 1
2
3
+704
genomic coordinate
B
5,345 epigenetic features
3 billion nucleotides
707 cell-type-specific transcription factor occupancy profiles
10⁷ variants
69
EUR GWAS summary statistics
42
EAS GWAS summary statistics
+
3x10⁹
TF binding presence or absence
ϵ {0,1}
observed data
p = P(Y = 1)
707
~
5,345
logistic regression model coefficients
ϵ ℝ
log p/(1 − p) ~ βX
X
3x10⁹
5,345
genome-wide epigenetic features ϵ {0,1}
log p/(1 − p) ~ βX
D
I. Heritability in shared regulatory elements
% h2 captured
EUR
EAS
1 2 3
1 2 3
annotations
II. High multi-ethnic marginal effect size correlation
(EAS lead SNPs)
f c g e h b
a i d
(EUR lead SNPs)
β̂_EAS
β̂_EUR
III. Trans-ethnic polygenic risk score models
IMPACT PRS: EUR PRS model with b, e, h
standard PRS: EUR PRS model with a, d, i
predict phenotype in EAS
accuracy
IMPACT PRS
standard PRS
trans-ethnic
within-pop
PRS improvement by IMPACT

Figure 3-1. Study design to identify regulatory annotations that prioritize regulatory variants in a multi-ethnic setting. A) Population-specific LD confounding and subsequent inflation of GWAS associations complicate the interpretation of summary statistics and transferability to other populations; functional data may help improve trans-ethnic genetic portability. B) Prism of functional data in IMPACT model: 707 genome-wide TF occupancy profiles (green), 5,345 genome-wide epigenomic feature profiles (blue), and fitted weights for these features (pink) to predict TF binding by logistic regression. Using IMPACT annotations, we investigate 111 GWAS summary datasets (yellow) of EUR and EAS origin. C) Compendium of 707 genome-wide cell-type-specific IMPACT regulatory annotations. D) Annotations that prioritize common regulatory variants must I) capture large proportions of heritability in both populations, II) account for consistent marginal effect size estimations between populations and III) improve the trans-ethnic application of PRS.

However, designing functional annotations that may improve PRS models is challenging. Functional annotations that best capture polygenic trait genetic variation must identify a large number of functional variants genome-wide without compromising specificity for trait-relevant regulatory programs. Pinpointing these mechanisms is especially difficult despite the fact that genome-wide association studies (GWAS) have identified thousands of genetic associations with complex phenotypes[18,51,52,97]. It has been estimated that about 90% of these associations reside in protein noncoding regions of the genome, making their mechanisms difficult to interpret[15,98]. Defining the etiology of complex traits and diseases requires knowledge of phenotyping-driving cell types in which these associated variants act. Transcription factors (TFs) are poised to orchestrate large polygenic regulatory programs as genetic variation in their target regions can modulate gene expression, often in cell-type-specific contexts[26,99]. Genomic annotations marking the precise location of TF-mediated cell type regulation can be exploited to elucidate the genetic basis of polygenic traits.

To overcome these challenges, we previously developed IMPACT, a genome-wide cell-type-specific regulatory annotation strategy that models the epigenetic pattern around TF binding using linear combinations of functional annotations[27]. In rheumatoid arthritis (RA), IMPACT CD4+ T cell annotations captured substantially more heritability than functional annotations derived from single experiments, including TF and histone modification ChIP-seq[24]. In this study, we expanded this approach by aggregating 5,345 functional annotations with an identical implementation of the IMPACT model framework using the same set of optimized parameters as previously calibrated. We created a powerful and generalizable resource of 707 cell-type-specific gene regulatory annotations (**Web Resources**) based on binding profiles of 142 TFs across 245 cell types (**Figure 3-1B,C**). This study builds on our previous work[27] in which we created 13 annotations (13 TF-cell type pairs) based on 515 functional annotations;

we observed remarkable consistency of IMPACT predictions for the same TF-cell type pair despite different training data and epigenetic features (**Figure B-1**). Assuming that causal variants are largely shared between populations[51,80], we hypothesized that restricting PRS models to variants within trait-relevant IMPACT annotations, which are more likely to have regulatory roles and less likely to be solely associated via linkage, will especially improve their trans-ethnic portability.

In this study, we identify key IMPACT regulatory annotations that capture genome-wide polygenic mechanisms underlying a diverse set of complex traits, supported by population non-specific enrichments of genetic heritability, multi-ethnic marginal effect size correlation (a possible mechanism of improved PRS), and improved trans-ethnic portability of PRS models (**Figure 3-1D**). Here, we defined and employed our compendium of 707 IMPACT regulatory annotations to study polygenic traits and diseases from 111 GWAS summary datasets of European (EUR) and East Asian (EAS) origin. Assuming shared causal variants between populations, annotations that prioritize shared regulatory variants must (1) capture disproportionately large amounts of genetic heritability in both populations, (2) be enriched for multi-ethnic marginal effect size correlation, and (3) improve the trans-ethnic applicability of population-specific PRS models. Using our compendium of regulatory annotations, we identified key annotations for each polygenic trait and demonstrated their utility in each of these three applications toward prioritization of shared regulatory variants. Overall, this work improves the interpretation and trans-ethnic portability of genetic data and provides implications for future clinical implementations of risk prediction models.

Material and Methods

Data

TF ChIP-seq data. On October 15, 2015, we downloaded all available transcription factor (TF) chromatin immunoprecipitation followed by sequencing (ChIP-seq) data derived from human primary cells or cell lines deposited on NCBI GEO (n = 13,732 datasets). Then we retained accessions for which input ChIP-seq (control data) were also generated and made publicly available (n = 3,181 of 13,732). We downloaded raw sequencing data in SRA format from NCBI GEO, then converted the data to FASTQ format using the SRA Toolkit function fastq-dump, used FastQC for quality assessment of sequencing reads, and finally mapped reads to the human genome (hg19/GRCh37) with Bowtie2 [v2.2.5] using default parameters. All ChIP-seq datasets were matched to corresponding control data from which peaks were called with macs [v2.1] with q value < 0.01 under a bimodal model, producing 3,181 bed file-formatted files[100,101]. For compatibility with the IMPACT method, we selected TFs with a known sequence motif, as recorded in the MEME database. Of the 442 TFs represented by the 3,181 TF ChIP-seq datasets, only 142 matched a known sequence motif, narrowing down the total number of considered datasets to 1,542. There was no dataset removal based on cell type classification. Of the 1,542 datasets (each characterized by a TF-cell type pair), there were 728 unique TF-cell type pairs, meaning many pairs have been assayed more than once. As described below in **Statistical Methods: *Training IMPACT***, we took the union of peaks among different experiments of the same TF-cell type pair. Therefore, the number of consolidated TF ChIP-seq datasets (n = 728 is < 1,542). Then for each of 728 datasets, we scanned TF ChIP-seq peaks for corresponding TF motifs, as described below in **Statistical Methods: *Training IMPACT***. We removed consolidated datasets with fewer than 7 peaks with TF motifs, the lower bound at which the logistic regression could converge, resulting in 707 consolidated datasets. Regarding the corresponding GEO accessions, this removal reduced the 1,542 utilized GEO accessions to 1,511. The 1,511 datasets account for 707 unique TF-cell type pairs, 142 unique TFs and 245 unique cell types or cell lines. These 1,511 datasets selected for use with our IMPACT model framework are described in **Table B-1**, including accession codes and experimental details.

Genome-wide annotation data. We augmented our set of 515 publicly available epigenomic and sequence feature annotations from our previous study[27] with 116 personally curated datasets from NCBI, 2,593 ENCODE histone ChIP-seq datasets and 2,121 ENCODE open chromatin DNase-seq datasets[102], all publicly available at the accessions provided in **Table B-2**. All files were collected in 6-column standard bed file format. This augmentation brought the total number of features to 5,345.

Genome-wide association data. We collected publicly available summary statistics data for 111 genome-wide association studies (GWAS) across separate cohorts of East Asian and European individuals[24,34,103]. East Asian GWAS data were collected from Biobank Japan (BBJ) while European GWAS data were collected from either UKBioBank (UKBB) or the GWAS catalog, referred to as PASS (publicly available summary statistics) (**Table B-3**). Since our analysis utilized S-LDSC which is based on the polygenic inheritance model, it is crucial to include summary statistics of GWAS conducted in large-scale samples[24]. First, we included summary statistics of EUR GWAS in which biologically plausible polygenic signals were confirmed in previous studies (**Table B-3**), beginning with the set of summary statistics (n = 42) we had previously downloaded from the Price Lab (**Web Resources**) and used in our previous work[27]. Next, we included additional diseases/traits for which both EAS (specifically BBJ) and EUR GWAS summary statistics are available. We chose to focus this study on EUR and EAS populations, as there is a very limited number of large GWAS in populations other than EUR and EAS[82,104,105]. As blood quantitative trait GWAS and disease GWAS were available from BBJ, we sought to collect matching EUR GWAS datasets to maximize phenotype overlap between populations. We included studies where cases were diagnosed by a physician and excluded studies which utilized self-reported cases, aiming to prepare comparable phenotypes between EAS and EUR GWAS. We downloaded such data from Riken, the Neale Lab, and the GWAS Catalog (**Web Resources**). In summary, we collected summary statistics of 42 EAS and 69 EUR GWAS. All summary statistics used had an observed scale heritability z-score > 1.96 as estimated by S-LDSC. All GWAS summary statistics were reformatted to be compatible with S-LDSC (see below) and thus contained the following information for each SNP (per row): rsID, A1 (reference allele), A2 (alternative allele), GWAS sample size (effective sample size per SNP, may vary with genotyping), chi-square statistic, *z*-score. For multi-ethnic genetic correlation and polygenic risk score prediction, all GWAS summary statistics were reformatted to contain the SNP ID (chr_position_A1_A2), chromosome, base pair, A1, A2, effect size estimate, effect size estimate standard error, and *P*-value.

Cell-type-specifically expressed gene set (SEG) and cell-type-specific histone modification (CTS) annotations. We downloaded 513 publicly available SEG annotations for European SNPs from phase 3 of 1000 Genomes (see **Web Resources**)[25]. SEG annotations are binary; each SNP is assigned a 1 or a 0, indicating that the SNP does or does not lie, respectively, within 100 kb of the gene body of the corresponding gene set[25]. We downloaded 220 publicly available CTS annotations of peak data in bed file format, from which we annotated European SNPs from phase 3 of 1000 Genomes[106] (see **Web Resources**)[24]. These annotations are also binary, in which case each SNP is designated a 1 or a 0, indicating that the SNP does or does not like, respectively, within the peak of histone modification. We also acquired the corresponding SEG and CTS SNP-level annotations for East Asian SNPs from phase 3 of 1000 Genomes from a previous study[103]. For all annotations, we used S-LDSC to compute LD scores and partitioned heritability using a customized version of the baselineLD annotations as described below.

Deep Learning annotations from DeepSEA and Basenji. For each commonly varying SNP, we assigned a sequence-mediated predicted activity score using two pre-trained deep learning models, DeepSEA[107] and Basenji[108]. We assigned two types of activity scores; 1) allelic-effect and 2) variant level, as per the nomenclature previously used[109]. For the DeepSEA model, the allelic-effect annotations represent the predicted change in the probability of TF binding, histone marks or DHS of the region around the SNP as a result of the change from reference to alternative allele. Similarly, for the Basenji model, the allelic-effect annotations represent the predicted change in aligned fragments to the region around the SNP as a result of the change from reference to alternative allele for DHSes, histone marks, or CAGE features. In both cases, we used pre-trained models from the respective studies with the recommended parameter settings used in the model training. These computations were performed using 1 GPU Tesla M40 card. For the allelic-effect activity score at a SNP, we take an ensemble of the predictions for the SNP over sequences with the SNP at the center, shifted 1 position to the left, or shifted 1 position to the right. For variant level predictions, we compared allelic-effect scores with the predicted epigenomic accessibility, characterized either by predicted number of aligned fragments for histone marks, DHS or CAGE features (as in Basenji) or predicted probability of TF binding, histone marks, or DHS (as in DeepSEA), in a 1 kb window around a SNP. These predictions are a denoised estimate of the Roadmap peak intensities as learned from sequence[109].

We downloaded 32 publicly available deep learning annotations for European SNPs from phase 3 of 1000 Genomes and used S-LDSC to compute LD scores (see **Web Resources**). The 32 annotations were comprised of Basenji[108] and DeepSEA[107] deep learning predictions corresponding to DHSes, H3K27ac, H3K4me1, and H3K4me3 meta-analyzed separately for blood and brain cell types and computed for both allelic effect and variant level models[109]. Additionally, we analyzed 78 new tissue-specific variant level and allelic effect annotations from DeepSEA and Basenji models. These 78 annotations corresponded to cell types that we identified as drivers of any of the five representative traits (asthma, height, MCV, RA, and PrCa). These 78 annotations extend beyond histone marks and DHS features used previously[109], accounting also for TF binding (DeepSEA) and CAGE features (Basenji). All 78 annotations are reported in **Table B-11**.

We also trained new allelic effect DeepSEA models on the TF ChIP-seq used to train what we identified as lead IMPACT annotations (13 unique) for the 21 traits investigated in the PRS analysis. We employed DeepSEA as previously described using default parameters, 1 Quadro GV100 (NVIDIA) GPU, Selene (v0.4.7), PyTorch (v1.3.1) [107,110]. For training the DeepSEA model, we used the genomic sequences corresponding to each of the 13 TF ChIP-seq peak sets as well as any regions where ENCODE or the Roadmap Epigenomics DeepSEA dataset contained at least one TF binding event. As done in the original DeepSEA study, we randomly sampled 1 kb sequences (hg19) from regions included ENCODE, Roadmap, or our TF ChIP-seq data. Considering each training TF ChIP-seq dataset separately, we determine positive samples as follows as done in the original DeepSEA study: if more than 100 bp of the center 200 bp of the 1kb sequence falls in our provided TF ChIP-seq peaks, this sequence is labeled with a 1, else 0. DeepSEA accurately predicted TF binding, average AUROC = 0.93, sem = 0.007; training was performed on chromosomes 1-5 and 10-22, testing was performed on chromosomes 8-9, and validation was performed on chromosomes 6-7.

BioBank Japan data. For PRS analysis, we utilized phenotype and genotype data of the BioBank Japan Project (BBJ)[111,112]. All of the calculations related to PRS were conducted on the RIKEN computing server. BBJ is a biobank that collaboratively collects DNA and serum samples from 12 medical institutions in Japan. This project recruited approximately 200,000 patients with the diagnosis of at least one of 47 diseases. Informed consent was obtained from all participants by following the protocols approved by their institutional ethical committees. We obtained approval from the ethics committees of the RIKEN Center for Integrative Medical Sciences and the Institute of Medical Sciences at the University of Tokyo.

Statistical Methods

IMPACT Model. We implemented our previously defined model to predict TF binding on a motif site. This model regresses the likelihood (*p*) of a binding event on the epigenomic profile of the motif site, in a logistic regression framework over *j* epigenomic features as follows:

$$log\,(\tfrac{p}{1-p}) \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j .$$

We use a weighted average of ridge and lasso regularization terms in the objective function to restrict the magnitude of fit coefficients and enforce sparsity to reduce overfitting, respectively, as follows:

$$argmin_{\beta} = (||Y - X\beta||^2 + \tfrac{1}{2}(1 - \alpha)||\beta||^2 + \alpha\, ||\beta||).$$

Training IMPACT. We trained an IMPACT model for each unique cell type-TF pair present in our data collection. Our collection consists of 3,181 TF ChIP-seq profiles, representing 442 TFs, 296 cell types, and 24 tissues. The IMPACT model requires that the assayed TF has a distinct binding motif and so we removed all ChIP-seq datasets corresponding to a TF that did not have a known sequence motif in MEME, Jaspar, or Transfac databases. This resulted in 1,542 TF ChIP-seq profiles across 142 TFs, 245 cell types, 23 tissues, and 728 unique combinations of TFs and cell types. As we did in our previous study[27], we merged experiments of the same TF-cell type combination by taking the union of the peaks. We next identified motif sites bound by a TF by using HOMER [v4.8.3][113] to scan ChIP-seq peaks for motif matches exceeding the empirically determined motif detection threshold. Similarly, we identified motif sites not bound by a TF by using HOMER to scan the entire genome for sequence matches. 21 of these models did not contain sufficient overlap between TF sequence motifs and ChIP-seq peaks which would lead to underfitting in the logistic regression (fewer than 7), thereby resulting in 707 total possible IMPACT annotations. We then trained 707 IMPACT models using up to 1,000 TF-bound sequence motifs (evidenced by ChIP-seq) and 10,000 unbound sequence motifs. To assess the predictive accuracy of IMPACT, we evaluated the AUPRC (area under the precision-recall curve) which is appropriate for classification tasks with considerable class imbalance. Accounting for the true ratio of bound to unbound motifs genome-wide, which is unique to each model but averages to 0.03, the average AUPRC was 0.53 (sem = 0.01). Using the class imbalance defined by the model (1,000 / 10,000 = 0.1), the average AUPRC was 0.74 (sem = 0.008). For each of 707 TF-cell type pairs, we learned a predictive model of TF binding and annotated SNPs genome-wide for both EUR and EAS populations, with a mean regulatory probability per nucleotide of 0.02 (sem = 7.5e-4).

Assessing cell type specificity of IMPACT tracks. We acquired lists of specifically expressed genes in 9 different cell types: T cells, B cells, fibroblasts, monocytes, brain, liver, colon, prostate, and breast according to differential gene expression *t*-statistics from previous work[25], specifically labeled as T.4+8int.Th, B.Fo.LN, Cells_Transformed_fibroblasts, Mo.6C+II-.LN, Brain_Cortex, Liver, Colon_Transverse, Prostate, Breast_Mammary_Tissue, respectively from either ImmGen or GTEx databases. Large and positive *t*-statistics represent greater specificity of gene expression in the target cell type, large but negative *t*-statistics represent specifically repressed genes, and *t*-statistics near 0 represent nonspecific gene expression, representing commonly expressed genes. For each cell type, we selected the 100 genes with highest *t*-statistics, e.g. specifically expressed (SE) genes, and 100 genes such that -0.5 < *t*-statistic < 0.5, e.g. not specifically expressed genes (NS). For each cell type separately, we collected all related IMPACT annotations from the compendium of 707 total annotations. Then for each annotation separately, we computed the average IMPACT score over all EUR SNPs from phase 3 of 1000 Genomes within 2kb of each SE or NS gene body. Finally, we computed the average across all 100 SE and 100 NS genes, separately.

Partitioning heritability with S-LDSC. We applied S-LDSC [v1.0.0][24] to partition the common (MAF > 5%) SNP heritability of 111 polygenic traits and diseases, with significantly non-zero heritability estimates ($P < 0.05$). Here, the term heritability is defined as previously[24], referring to inferences made by S-LDSC about heritability causally explained by common SNPs. This is a different quantity than genotyping-array-based SNP-heritability[114,115]. We partitioned heritability using a customized version of the baselineLD model, in which we excluded cell-type-specific regulatory annotations (as we would be testing the enrichment of such annotations from IMPACT). In total, we used 69 cell-type-nonspecific baselineLD annotations and added one or more IMPACT annotations to the model to test for cell-type-specific enrichment. We use three metrics to evaluate how well our IMPACT annotations capture polygenic heritability: enrichment[24], the proportion of heritability explained by the top 5% of SNPs[24], and per-annotation standardized effect size, τ *[34]. Briefly, enrichment is defined as the proportion of common SNP heritability divided by the genome-wide proportion of SNPs in the annotation, for continuous annotations this is the average annotation value across SNPs. τ * represents the average per-SNP heritability of a category of SNPs, where a single SNP may claim membership to one or more categories. τ * is defined as the proportionate change in per-SNP heritability associated with a one standard deviation increase in the value of the annotation. The sum of the τ * over categories of SNPs equals the total estimated heritability of the trait. τ * has units of heritability and is comparable between traits, annotations, and populations, because it is normalized for the total heritability (indicative of the power of the GWAS), the dispersion of the annotation values (annotation size), and the number of common SNPs (population-specific) considered in the model, respectively. τ , the precursor of τ *, is the coefficient estimated in the S-LDSC regression. τ and τ * are conditionally dependent on the provided baselineLD annotations. Therefore, the τ * estimate for an IMPACT annotation is considered a measure of cell-type-specific or annotation-specific SNP heritability, as the remaining annotations in the model (baselineLD) are not cell-type-specific. Significance of τ * is computed using a *z*-test of how different the τ^* estimate is from 0; the significance of strictly positive τ^* estimates are reported in our study. A negative τ^* would indicate a depletion of heritability, suggesting that lower values of the annotation are more enriched for trait-associated genetic variation.

Measuring heritability in top X% of SNPs of a continuous annotation. To partition the heritability captured by various top echelons of SNPs of a given continuous annotation, we used the same strategy as in a previous study[34]. By this strategy, the proportion of heritability explained by a set of SNPs is the sum over all SNPs of the product of the τ^* of each category in the S-LDSC model, e.g. baselineLD plus IMPACT annotation, and the SNP membership to that category (1 or 0 in the case of binary annotations, continuous values in the case of continuous annotations) divided by the same metric for all SNPs genome-wide.

Conditional S-LDSC analysis to identify independent annotation-trait associations. Due to the redundancy in modeled cell type programs and inherent covariance of IMPACT annotations (**Figure B-3**), the τ^* associations we find with S-LDSC cannot be independent. To this end, for each of 95 traits across EUR and EAS for which we identified a lead IMPACT annotation, reported in **Table B-9**, we performed a series of conditional analyses using S-LDSC. For each trait with more than one significant τ^* association, we created S-LDSC models consisting of the 69 baselineLD annotations, the lead annotation for that trait, and separately, each remaining significant IMPACT annotation. We kept annotations that retained their τ^* significance when conditioned on the lead annotation(s), which we also required to retain significance. We iteratively performed these conditional analyses until we were no longer able to identify independent τ^* associations.

Deming regression of EUR τ * on EAS τ *. As there is significant correlation among IMPACT annotations, due to redundancy in cell type regulatory elements, we used an iterative pruning approach, similar to LD-pruning, to identify independent IMPACT annotations. For each trait, we ranked all 707 IMPACT annotations by their τ * significance values. Then, we selected the lead annotation, removed all annotations correlated with Pearson $r > 0.5$, and selected the next lead annotation, and so on. This approach produced a set of relatively independent annotations, for which the assumptions of Deming, or any, regression would not be violated. For each trait, we ran Deming regression over approximately 100 independent IMPACT annotations using the R function *deming* within the package *deming*. Across independent observations for all traits, we tested the null hypothesis that the slope of the Deming regression, which considers standard errors on both the predictor (EUR τ *) and response variables (EAS τ *), is equal to 1.

Multi-ethnic and within-population genetic correlation. We computed the genetic correlation (R_g) between pairs of 29 traits for which we acquired EUR and EAS GWAS using Popcorn [v.0.9.6][116] with default parameters, including maximum likelihood estimation as opposed to regression[81]. First, we computed cross-population scores between the two populations using the *compute* flag with the *popcorn* executable, indicating approximately the correlation between LD at each SNP using EUR and EAS reference LD panels from phase 3 of 1000 Genomes. Then, we used the *fit* flag with the *popcorn* executable to compute the multi-ethnic genetic correlation of these 29 traits. R_g estimates computed after restricting to MAF > 5% did not significantly differ from no MAF restriction. Popcorn computes R_g using either "genetic impact" (effect sizes normalized by allele frequency) or "genetic effect" (unmodified effect sizes). We observed no significant heterogeneity between the R_g computed using "genetic impact" and "effect", although "genetic effect" estimates were consistently but not significantly larger.

We then computed cross-trait cross-population genetic correlations across 21 traits for which we observed at least one significant IMPACT annotation association in both EUR and EAS. Therefore, in total we computed the genetic correlation among 42 traits (21 phenotypes x 2 populations). For pairs of traits with one from EUR and one from EAS, we used Popcorn as described above with MAF threshold of 5% and “genetic impact”. For pairs of traits from the same population we used LDSC [v.1.0.0]. First we used the *munge_sumstats.py* script to make the direction of allelic effect consistent in the GWAS summary statistics while also restricting to well-imputed Hapmap3 SNPs. Then, we used the *ldsc.py* script with the *-rg* flag to compute the genetic correlation using EUR and EAS reference LD panels from phase 3 of 1000 Genomes where appropriate.

Multi-ethnic marginal effect size correlation, heterozygosity correlation, and F_{st} **.** We acquired GWAS summary statistics for each of 21 shared traits between EUR and EAS for which there was at least one significant IMPACT association in each population. Then, we restricted to SNPs shared between EUR and EAS GWAS summary statistics. Next, we performed stringent iterative LD clumping with PLINK [v1.90b3][117] using EUR summary statistics (selecting the most significant SNP, then removing all SNPs in LD with r^2 > 0.1 within 1 Mb, then selecting the next most significant SNP, and so on). This step satisfies the assumption of independence in the Pearson correlation that we will compute among marginal effect sizes. We selected our initial set of SNPs under three scenarios: (1) using no functional inference, (2) using the top 5% of SNPs according to the trait’s lead EUR IMPACT annotation, and (3) using the bottom 95% of SNPs according to the trait’s lead EUR IMPACT annotation (mutually exclusive with scenario 2). With our set of independent SNPs for each trait and under each of three scenarios, we compute a Pearson correlation between the estimated effect sizes, while further stratifying loci on 17 EUR *P*-values (1, 0.3, 0.1, 0.03, 0.01, 3e-3, 1e-3, 3e-4, 1e-4, 3e-5, 1e-5, 3e-6, 1e-6, 3e-7, 1e-7, 3e-8, 1e-8). For example, stratum with *P* = 0.1 includes all SNPs with EUR GWAS *P* < 0.1. Similarly, we computed the Pearson correlation of the EUR and EAS heterozygosity, defined as 2pq, where p is the reference allele frequency and q is the alternative allele frequency, using the same sets of variants as described above. Furthermore, we computed the F_{st} , where large values indicate a reduction in heterozygosity, at each variant and average F_{st} for each set of variants at each *P* value threshold for each of 21 considered traits. To this end, we collected the alternative allele frequencies from 1000G for EUR (EUR_{AF}) and EAS (EAS_{AF}) populations and defined F_{st} as the following:

$$F_{st} = (EUR_{AF} - EAS_{AF})^2 / (2p(1-p)),$$

where *p* is the average between EUR_{AF} and EAS_{AF} .

Polygenic risk score calculation. In this study, we utilized pruning and thresholding (P+T) for the calculation of PRS. We constructed PRS models from either EUR summary statistics or EAS summary statistics and evaluated their predictive performance on individual EAS phenotypes. Here, we define within-population PRS as PRSEAS and trans-ethnic PRS as PRSEUR to avoid confusion. For PRSEUR, we utilized genome-wide summary statistics from EUR as reported in their publicly available version. For PRSEAS, we held out 5,000 individuals for PRS analysis and conducted GWAS using the remaining individuals to avoid overfitting (see next section). For each trait separately, we restricted our analysis to variants that exist in both GWAS summary statistics and post-imputation genotype data of EAS individuals used for PRS analysis (imputation quality of r^2 > 0.3 in minimac3). A detailed description related to the genotyping platform and imputation strategy is provided in a previous report[101]. We excluded the MHC region in this analysis.

We designed PRS models using two strategies: standard PRS and functionally-informed PRS. For standard PRS$_{EUR}$, we performed conventional LD clumping to acquire sets of independent SNPs using EUR LD reference panels from phase3 of 1000 Genomes. Similarly for PRS$_{EAS}$, we utilized EAS LD reference panels from phase3 of 1000 Genomes. We used PLINK [v1.90b3][117] to remove variants in LD with r^2 > 0.2 with a significance threshold for index SNPs of P = 0.5. For functionally-informed PRS, we restricted the analysis to variants with high IMPACT score according to the lead IMPACT annotation before conducting LD clumping. As before, we define the lead annotation as the one with the largest τ * estimate that was significantly greater than 0. When we designed PRS$_{EUR}$, we utilized the lead IMPACT annotation in EUR GWAS summary statistics (EAS summary statistics were not taken into account to avoid overfitting). Similarly, when we design PRS$_{EUR}$, we utilized the lead IMPACT annotation in EAS GWAS summary statistics for which 5,000 EAS individuals for PRS analysis were removed to avoid overfitting. We performed LD clumping using variants within a predefined top percentage of IMPACT scores. This was determined by the percentage that captured the closest to 50% of total trait heritability; considered percentages included the top 1%, 5%, 10%, and 50%.

We evaluated PRS performance using EAS individuals. First, we used all individuals in the BBJ cohort for PRS$_{EUR}$ testing. Second, we compared the improvement afforded by IMPACT in PRS$_{EUR}$ relative to PRS$_{EAS}$ models using 5,000 randomly selected individuals in BBJ; specifically for case-control GWAS, we randomly selected 1,000 cases and 4,000 controls.

For all models, we built a PRS for each individual j in our test set (in all cases, there is no overlap between GWAS samples and PRS samples) using variant effect size estimates from GWAS as follows:

$$PRS_j = \sum_{i}^{M} A_{j,i} * \beta_i , \qquad \text{(Equation 1)}$$

Where M is the total number of SNPs shared between GWAS summary statistics and post-imputation genotype data of EAS individuals, *i* is the i^{th} SNP in the model, $A_{j,i}$ is the allelic dosage of the trait-increasing allele *i* in individual *j*, and β_i is the estimated effect size of allele *i* from GWAS. We calculated PRS using PLINK2.

For QC of quantitative phenotypes, we excluded (1) related samples (PI_HAT > 0.187 estimated by PLINK), (2) samples with age < 18 and age > 85, and (3) samples with measured values outside three interquartile ranges (IQR) of the upper or lower quartiles. The effect of sex, age, age^2 , the top 10 PCs, and affection status of 47 diseases were removed by linear regression, and the residuals were further normalized by the rank-based inverse normal transformation (see Equation 3 below). For QC of case/control phenotypes, we excluded (1) related samples (PI_HAT > 0.187 estimated by PLINK) and (2) samples with age < 18 and age > 85.

We then regressed our phenotype of interest (Y), a measured quantitative trait or a diagnosed disease among the PRS samples, on the per-individual PRS as follows:

For diseases,

$$Y_j \sim PRS_j + sex + age + Geno\,PC1 + ... + Geno\,PC10. \qquad \text{(Equation 2)}$$

For quantitative traits,

$$Normalized\,Y_j \sim PRS_j. \qquad \text{(Equation 3)}$$

We then report the variance explained; for quantitative traits, this is the variance explained by a linear model and for diseases, the variance explained is from a logistic model (Nagelkerke R^2)[80,81,118] which we convert to liability scale pseudo R^2 such that R^2 values are comparable among both quantitative and case/control phenotypes. We used various GWAS *P* value thresholds (0.1, 0.03, 0.01, 0.003, 0.001, 3e-4, 1e-4, 3e-5, 1e-5) to assess the predictive performance of our PRS. For each model, we reported in the text the largest R^2 achieved across the nine P value thresholds. For case/control traits, while R^2 estimates are reported on the liability scale, effect size estimates were derived on the logistic scale. To ensure the robustness of our results to the scale on which effect sizes are estimated, we converted logistic β to probit and then to liability scale, using this previously published conversion[119]. For EAS traits, the disease prevalence required for conversion from logistic to probit was derived from the Japanese epidemiological census[120] and for EUR traits, the prevalences were derived from previous studies: for asthma[121], for RA[122], for PrCa[123], for CAD[124], and for T2D[125]. The allele frequencies required for conversion from probit to liability were derived from 1000 Genomes of the corresponding population.

To estimate confidence intervals of PRS performance (R^2 , as explained above), we conducted 1,000 bootstraps using the R package *boot*. We also conducted 10,000 bootstraps to evaluate whether the R^2 difference between two PRS models (functionally-informed - standard) is significantly greater than 0; we calculated the R^2 difference between two PRS models in each round of bootstrapping (delta R^2), and assess its distribution in 10,000 bootstraps. If we let N be the frequency of delta $R^2 < 0$, we define one-tailed *P* values for delta $R^2 > 0$ as (N + 1)/10,000. We also estimated confidence intervals of PRS performance using a block jackknife across the genome as previously done[23], using 200 adjacent genomic bins of equal size. Then iteratively, one bin of variants was removed from the PRS model and the R^2 estimate was recalculated to establish a confidence interval around the original estimate. We additionally estimated confidence intervals around the difference between IMPACT PRS R^2 and standard P+T R^2 using a block jackknife.

Genome-wide association studies in BBJ. As described in the previous section, we held out 5,000 randomly selected individuals for the PRS analysis and performed GWAS on the remaining individuals (sample sizes are provided in **ST16-17**). GWAS was conducted with PLINK2 using the same imputed dosages as used in the PRS analysis. For quantitative traits, normalized residuals were analyzed by a linear regression model. For diseases, affection status was analyzed by a logistic regression model using age, sex, and the top 10 genotype PCs as covariates.

PRS distributions in 1000G subpopulations. To address if there was any global bias in PRS distributions that IMPACT variant prioritization could mitigate, we computed PRS based on EUR and EAS summary statistics as done above and allelic dosages of five different 1000 Genomes populations (AFR, AMR, EAS, EUR, and SAS). Then we used anova to compute the F-statistic indicative of the inter-population variance and compared PRS with IMPACT prioritization to those with no variant prioritization.

Results

Building a compendium of *in silico* gene regulatory annotations

To capture genetic heritability of diverse polygenic diseases and quantitative traits, we constructed a comprehensive compendium of 707 cell type regulatory annotation tracks. To do this, we applied the IMPACT[27] framework to 707 unique TF-cell type pairs obtained from a total of 3,181 TF ChIP-seq datasets from NCBI, representing 245 cell types and 142 TFs with known sequence motifs (**Figure 3-1B**, **Material and Methods**, **Web Resources**, **Table B-1**, **Figure B-2**)[100]. We provide publicly available open-source software (see **Web Resources**) corresponding to the analyses presented in this manuscript. We caution that the 707 TF/cell type pairs represented in publicly available data is a small fraction of the total possible pairs of 142 TFs and 245 cell types (n = 34,790), although there are several experimental and practical reasons why this theoretical maximum is not reached (**Discussion**). Briefly, IMPACT learns an epigenetic signature of active TF binding evidenced by ChIP-seq, differentiating bound from unbound TF sequence motifs using logistic regression. We derive this signature from 5,345 epigenetic and sequence features, predominantly generated by ENCODE[102] and Roadmap[126] (**Material and Methods**, **Table B-2**); these data were drawn from diverse cell types, representing the biological range of the 707 candidate models. IMPACT then probabilistically annotates the genome, e.g. on a scale from 0 to 1, without using the TF motif, identifying regulatory regions that are similar to those that the TF binds.

To assess the specificity of our IMPACT annotations, we test whether they (1) accurately predict binding of the modeled TF, (2) share cell-type-specific characteristics with other tracks of the same cell type, and (3) score cell-type-specifically expressed genes higher than nonspecific genes. The 707 models that we defined had a high TF binding prediction accuracy with mean AUPRC = 0.54 (sem = 0.01, **Material and Methods**, **Figure B-3**) using cross-validation.

Annotations segregated by cell type rather than by TF, excluding CTCF, suggesting the same TF may bind to different enhancers in different cell types (**Figure 3-2A**). On average, we observed that annotations of the same cell types were more strongly correlated genome-wide (Pearson *r* = 0.56, sem = 0.02) than annotations of different cell types (Pearson *r* = 0.48, sem = 0.01, one-tailed difference of means *P* < 0.001, **Figure B-3**). Furthermore, the covariance structure between TF ChIP-seq training datasets is similar to that of corresponding IMPACT annotations, indicating that the IMPACT model does not introduce spurious correlations among unrelated ChIP-seq datasets (**Figure B-3**). Lastly, for nine different cell types, we examined cell-type-specifically expressed genes from Finucane et al[25] and corresponding differential expression *t*-statistics. For each of nine cell types, we observed larger cell-type-specific IMPACT probabilities at SNPs in and near cell-type-specific genes compared to generally expressed genes (mean fold-change across 10 to 99 cell-type-specific IMPACT tracks ranged from 1.08 to 1.96 across nine cell types, one-tailed paired wilcoxon *P* < 0.04 for seven of nine cell types, **Figure 3-2B**, **Figure B-3**, **Material and Methods**), suggesting that IMPACT annotates relevant cell type regulatory elements.

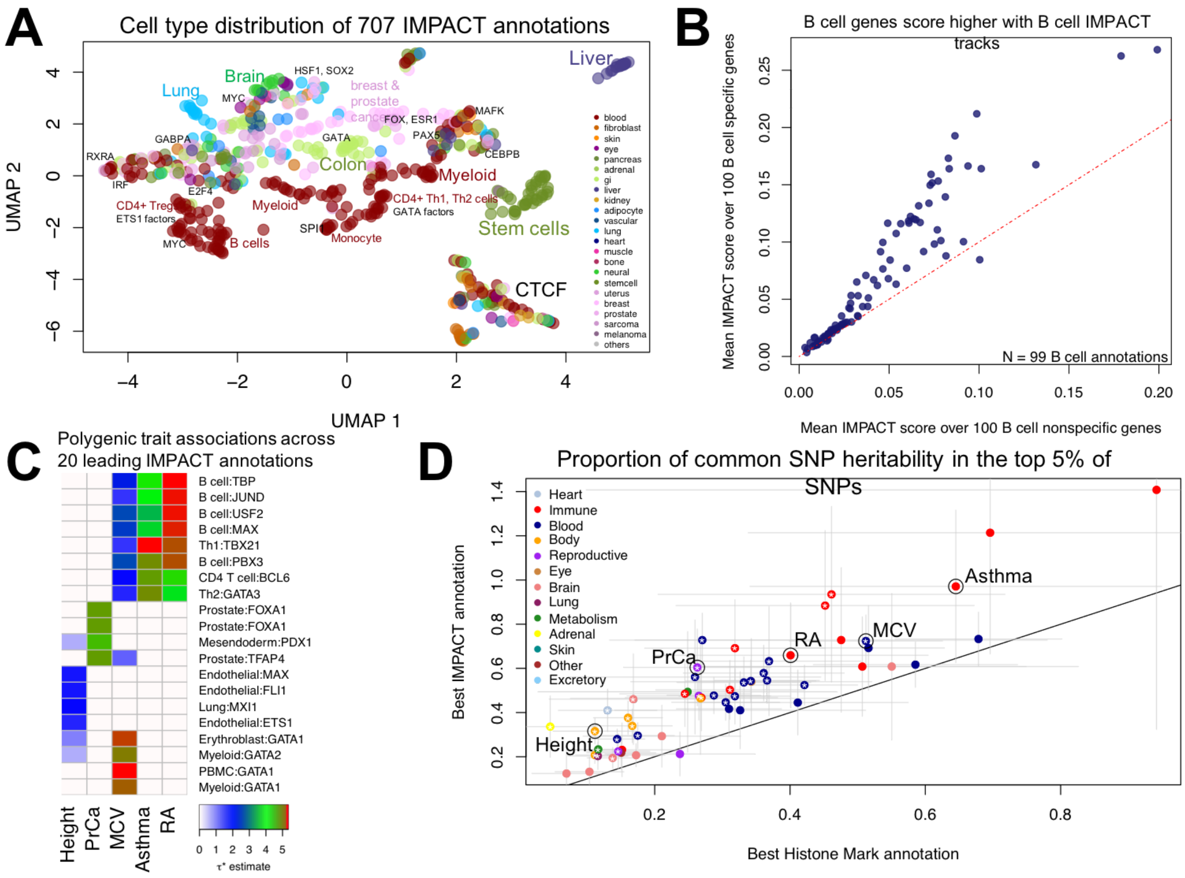


Figure 3-2. IMPACT annotates relevant cell type regulatory elements. A) Low-dimensional embedding and clustering of 707 IMPACT annotations using uniform manifold approximation projection (UMAP). Annotations colored by cell type category; TF groups indicated where applicable. B) IMPACT annotates cell type specifically expressed genes with higher scores than nonspecific genes. C) Biologically distinct regulatory modules revealed by cell type-trait associations with significantly nonzero τ *. Shown here are the 5 representative EUR complex traits and the 4 leading IMPACT annotations for each, resulting in 20 IMPACT annotations highlighted from 707 total. Color indicates τ * value. D) Lead IMPACT annotations capture more heritability than lead cell-type-specific histone modifications across 60 of 69 EUR summary statistics for which a lead IMPACT annotation was identified. τ * indicates heritability estimate difference of means *P* < 0.05. Gray segments indicate the 95% CI around the heritability estimate.

Partitioning common SNP heritability of 111 GWAS summary statistics in EUR and EAS

We obtained summary statistics from 111 publicly available GWAS for diverse polygenic traits and diseases. For narrative purposes throughout the text, we use five genetically uncorrelated (R_g point estimates between traits ranged from -0.08 to 0.20, **Table B-3**, although no R_g was significantly different from 0, all two-tailed z test *P* > 0.40 after Bonferroni correction for 10 pairs) and biologically diverse traits that capture the spectrum of summary statistics analyzed in order to exemplify our results in addition to reporting metrics averaged over all traits analyzed. These five traits include an allergic phenotype: asthma, an autoimmune disease: RA, a neoplastic type: prostate cancer (PrCa), a hematological quantitative trait: mean corpuscular volume (MCV), and an anthropometric trait: height. These included 69 from EUR participants[27,34] (average N = 180K, average heritability z-score = 12.9, 41/69 from UK BioBank)[24,127] and 42 from EAS participants of BioBank Japan[81,101,128,129] (average N = 157K, average heritability z-score = 6.6)[52] (**Table B-3**). We chose to focus our study on EUR and EAS populations, as there is a limited number of large GWAS in populations other than EUR and EAS[82,104,105]. All of the summary statistics used were generated from studies that had a sample size greater than 10,000 individuals and also had a significantly non-zero heritability (z-score > 1.97). There are 29 phenotypes for which we obtained summary statistics in both EUR and EAS. We were interested to see if any traits had a multi-ethnic genetic correlation that deviated from 1. Therefore, we explicitly tested this and found that 16 traits have multi-ethnic R_g that does not deviate from 1 (one-tailed z test *P* > 0.05/29 tested traits), while 13 traits have multi-ethnic R_g that does deviate from 1 (one-tailed z test *P* < 0.05/29 tested traits). Overall we observed high R_g for most traits, supporting our assumption that causal variants are generally shared across populations (**Material and Methods**, **Figure B-4**)[130]. At two extremes, basophil count has a low multi-ethnic R_g of 0.32 (sd = 0.10), while atrial fibrillation has a high multi-ethnic R_g of 0.98 (sd = 0.11), consistent with previous observations made using *Popcorn*, but using different parameter estimation strategies (**Material and Methods**)[81].

We then partitioned the common SNP (minor allele frequency (MAF) > 5%) heritability of these 111 datasets using S-LDSC[24] with an adapted baseline-LD model excluding cell-type-specific annotations[27,34] (**Figure B-4, Material and Methods**). Here, heritability refers to the inferences made by S-LDSC about the heritability causally explained by common SNPs as defined previously[24], as opposed to genotyping-array-based SNP-heritability[114,115] or other definitions. We caution that the results presented herein are a consequence of the analyzed GWAS populations, polygenic traits and diseases, and available experimental data to create functional annotations. Next, we tested each of the traits against each of the 707 IMPACT annotations, assessing the significance of a non-zero τ *, which is defined as the proportionate change in per-SNP heritability associated with a one standard deviation increase in the value of the annotation (**Material and Methods**)[34]. Of 707 by 111 (n = 78,477) possible associations subjected to 5% FDR, we detected 7,993 associations, 5% of which we expect to be false positives. We observed that 95 phenotypes had at least one significant annotation-trait association (τ * > 0, two-tailed z test *P* < 0.05 at 5% FDR, **Ext. Data B-1, Material and Methods**, **Tables B-4-8**). Here, we highlight the four leading IMPACT annotations associated with EUR summary statistics for each of the five exemplary phenotypes mentioned above: asthma, RA, PrCa, MCV, and height (**Figure 3-2C**, associations between all traits and annotations in **Ext. Data B-1**). Consistent with known biology, B and T cells were strongly associated with asthma[131], RA[1], and MCV[132,133] while other blood cell regulatory annotations predominantly derived from GATA factors were also associated with MCV. Prostate cancer cell lines were associated with PrCa, while many cell types including myoblasts[134], fibroblasts[135], and adipocytes[136,137], lung cells, and endothelial cells were associated with height, perhaps related to musculo-skeletal developmental pathways.

For each trait, we defined the lead IMPACT regulatory annotation as the annotation capturing the greatest per-SNP heritability, e.g. the largest, while significant, τ^* estimate (**Table B-9**). With the top 5% of SNPs, lead IMPACT annotations captured an average of 43.3% of common SNP heritability (sem = 2.8%) across these 95 polygenic traits (**Figure B-5**, **Material and Methods**), with more than 25% of heritability captured for two-thirds of the tested summary statistics (73/111 traits) and more than 50% captured for 28% (31/111). Identifying functional annotations that capture large proportions of heritability is an important step to understanding biological mechanisms of genetic variation. We observed higher heritability enrichments for autoimmune diseases and hematological traits, likely due to the abundance of blood cell types represented by our IMPACT annotations and possibly due to a single or a few related causal cell types. On the other hand, we observed lower heritability enrichment for brain-related, lung-related, and adrenal traits, likely due to the underrepresentation of relevant tissue or cell types in the TF ChIP-seq data and possibly due to multiple different causal cell types. We observed significantly greater τ^* of lead IMPACT annotations among traits with lower estimated polygenicity (linear regression coefficient = -0.11, $P < 3.97e\text{-}5$). Traits with higher polygenicity may be driven by more than one causal cell type; therefore a single IMPACT annotation may capture a smaller proportion of total common SNP heritability. Returning to our five exemplary phenotypes, with the top 5% of EUR SNPs, IMPACT captured 97.1% (sd = 17.6%) of asthma heritability with the T-bet Th1 annotation, 65.9% (sd = 12.1%) of RA heritability with the B cell TBP annotation, 60.4% (sd = 8.9%) of PrCa heritability with the prostate cancer cell line (LNCAP) TFAP4 annotation, 72.4% (sd = 6.0%) of MCV heritability with the GATA1 PBMC annotation, and lastly 31.6% (sd = 3.0%) of height heritability with the lung MXI1 annotation

(**Figure 3-2D**). While the observed association between lung and height is not intuitive, within the MXI1 gene lies a genome-wide significant variant associated with height[138].

To demonstrate the value of IMPACT tracks, we compared them to annotations derived from single experimental assays and from machine learning models. For example, since each of the IMPACT tracks was trained on TF ChIP-seq data, we compared the per-annotation standardized effect sizes (τ *) achieved by both annotation types. We observed that on average the τ * of lead IMPACT annotations (mean τ * = 3.53, sem = 0.91) was greater than by the analogous TF ChIP-seq used in training (mean τ * = 1.71, sem = 0.94, across 95 traits one-tailed paired wilcoxon *P* < 2.6e-16). We then compared IMPACT tracks to histone marks, which are commonly used to quantify cell type heritability[24]. From 220 publicly available cell-type-specific histone mark ChIP-seq annotations of EUR SNPs[24], we selected the lead histone mark track for each of 69 EUR summary statistics (**Material and Methods**). Restricting to the top 5% of SNPs, we observed that the mean EUR heritability captured by lead IMPACT annotations (49.5%, sem = 3.2%) was on average greater than by lead histone mark annotations (29.1%, sem = 2.5%, one-tailed paired wilcoxon *P* < 8.8e-12, **Figure 3-2D, Table B-10**). For example, the lead IMPACT annotation for asthma captured 64.2% (sd = 15.5%) of heritability, 1.5x more heritability than the lead histone mark annotation (H3K27ac in CD4+ Th2). Similarly, IMPACT captured 1.7x more RA heritability than H3K4me3 in CD4+ Th17s; IMPACT captured 1.4x more MCV heritability than H3K4me3 in CD34+ cells; IMPACT captured 2.3x more PrCa heritability than H3K4me3 in CD34+ cells; and IMPACT captured 3.1x more height heritability than H3K4me3 in lung cells. In terms of τ *, IMPACT also captured more per-SNP heritability than histone marks (one-tailed paired wilcoxon *P* < 9.1e-9, mean τ * fold change across traits = 1.38x, **Figure B-6**). We further compared the heritability captured by IMPACT to annotations created from state-of-the-art deep learning algorithms trained to predict various regulatory element marks, Basenji[108] and DeepSEA[107]. Performing a comprehensive analysis is challenging for two reasons. First, there is a limited set of genome-wide SNP-level deep learning predictions in the public domain with the exception of a few studies[109]. Second, as deep learning models are specific to a particular functional mark, comprehensive genome-wide cataloging is a combinatorially large problem which grows with the number of tested cell types, functional marks, and model types. Therefore, we performed the most comprehensive analysis that was feasible, focusing on the five representative traits. To this end, we collected 123 relevant deep learning annotations to target these traits (**Table B-11, Material and Methods**) and selected the lead deep learning track for each trait (**Material and Methods**). We observed that for each of five traits, the lead IMPACT annotation generally captured more heritability in the top 5% of SNPs (mean = 65.4%, sem = 10.9%) and resulted in generally larger τ * (mean = 4.4, sem = 0.70) than the lead deep learning annotations (heritability mean = 39.1%, sem = 1.9%, τ * mean = 1.6, sem = 0.30, one-tailed paired wilcoxon *P* = 0.031 for both heritability and τ *, **Figure B-7**). Although limited by the availability of deep learning annotations, we further compared lead IMPACT annotations to lead deep learning annotations across all 69 EUR traits and in all cases IMPACT trended toward higher heritability and τ * (Basenji heritability comparison one-tailed paired wilcoxon *P* < 2.0e-11, DeepSEA heritability comparison *P* < 1.4e-10, Basenji τ * comparison one-tailed paired wilcoxon *P* < 3.4e-11, DeepSEA τ * comparison *P* < 8.8e-12, **Appendix B**, **Figure B-8**, **Table B-13**).

Since some of our IMPACT annotations are similar to each other (**Figure B-3**), we performed serial conditional analyses in order to identify IMPACT annotations explaining heritability independently from one another (**Material and Methods**). This strategy might identify complex traits for which several distinct biological mechanisms are independently regulated by genetic variation. Indeed, we identified 30 EUR phenotypes and 8 EAS phenotypes with multiple independent IMPACT associations (**Figure B-9**, **Table B-14-15**). For example, four IMPACT annotations were independently associated with EUR PrCa: prostate (TFAP4), prostate (RUNX2), mesendoderm (PDX1), and cervix (NFYB). Moreover, for seven EUR traits, three IMPACT annotations were independently associated: height (adipocytes, fibroblasts, lung), neutrophil count (monocytes, adipocytes, B cells), osteoporosis (myoblasts, mesenchymal stem cells, cervix), IBD (T cells and two B cell annotations), platelet count (PBMCs, hematopoietic progenitors, muscle), systolic blood pressure (endothelial, mesenchymal stem cells, fibroblasts), and white blood cell count (B cells, adipocytes, hematopoietic progenitors). Among functionally correlated traits, we observed consistency in the independently associated IMPACT annotations, proposing a biological basis for genetic correlation (**Appendix B**). In general, identifying functional concordance among traits with genetic correlation less than 1 provides a quantitative biological basis for the dissimilarity between traits that is orthogonal to genetic correlation approaches[130,139–142]. We found that the heritability z-score, an index correlated with the power of S-LDSC[24], is strongly predictive of the number of independent regulatory associations (linear regression coefficient = 0.06, $P < 1.2e\text{-}5$), while sample size is not (linear regression $P = 0.59$) (**Figure B-10**). Our findings suggest that multiple independent regulatory programs can contribute to the heritability of complex traits, and we can detect them when phenotypes are sufficiently heritable and the GWAS provide accurate effect size estimation.

Concordance of polygenic regulation between European and East Asian populations

Previous studies have shown concordance of polygenic effects between EUR and EAS individuals in RA[79] and between EUR and African American individuals in PrCa[143]. However, to our knowledge, the extent of these shared effects has not yet been comprehensively investigated across many functional annotations and in diverse traits. Assuming shared causal variants in EUR and EAS, IMPACT annotations that best prioritize shared genomic regions regulating a phenotype presumably also disproportionately capture similar amounts of heritability in both EUR and EAS (**Figure 3-1D-I**, **Figure 3-3A**). Here, we quantified the SNP heritability (τ *) of 29 traits in EUR and EAS captured by a set of approximately 100 independent IMPACT regulatory annotations (**Figure 3-3B, Figure B-11, Material and Methods**). Briefly, we selected independent annotations using an iterative pruning approach: for each trait, we ranked all annotations by τ * and removed any annotation correlated with Pearson r^2 > 0.5 to the lead annotation and then repeated. As IMPACT annotations are independent of population-specific factors including LD and allele frequencies (**Figure B-4**), they are poised to capture the genome-wide distribution of regulatory variation in a population-independent manner. We observed that τ * estimates across annotations for EUR and EAS are strikingly similar, with a regression coefficient that is consistent with identity (slope = 0.98, sem = 0.04). For example, we observed a strong Pearson correlation of τ * between EUR and EAS for asthma (*r* = 0.98), RA (*r* = 0.87), MCV (*r* = 0.96), PrCa (*r* = 0.90), and height (*r* = 0.96). Cross-ancestry functional concordance is not specific to IMPACT annotations as we observed a similar relationship among cell-type-specific histone marks using the same strategy (**Figure B-12**)[52]. Additionally considering 513 cell-type-specifically expressed gene sets (SEG)[25,52], we could not observe cross-ancestry concordance due to too few significant associations shared between populations. Furthermore, we found that none of our τ * estimates show evidence of population heterogeneity (all two-tailed difference of means *P* > 0.56 at 5% FDR). This might be a result of noise around the τ * estimates, such that true heterogeneity is too subtle to detect in this regime. Overall, our results suggest that regulatory variants in EUR and EAS populations are equally enriched within the same classes of regulatory elements. This does not exclude the possibility of population-specific variants or causal effect sizes, as evidenced by 13 traits with multi-ethnic genetic correlation significantly less than 1 (*P* < 0.05/29 tested traits). Rather, these results suggest that causal biology, including disease-driving cell types and their regulatory elements, underlying polygenic traits and diseases, is largely shared between these populations.

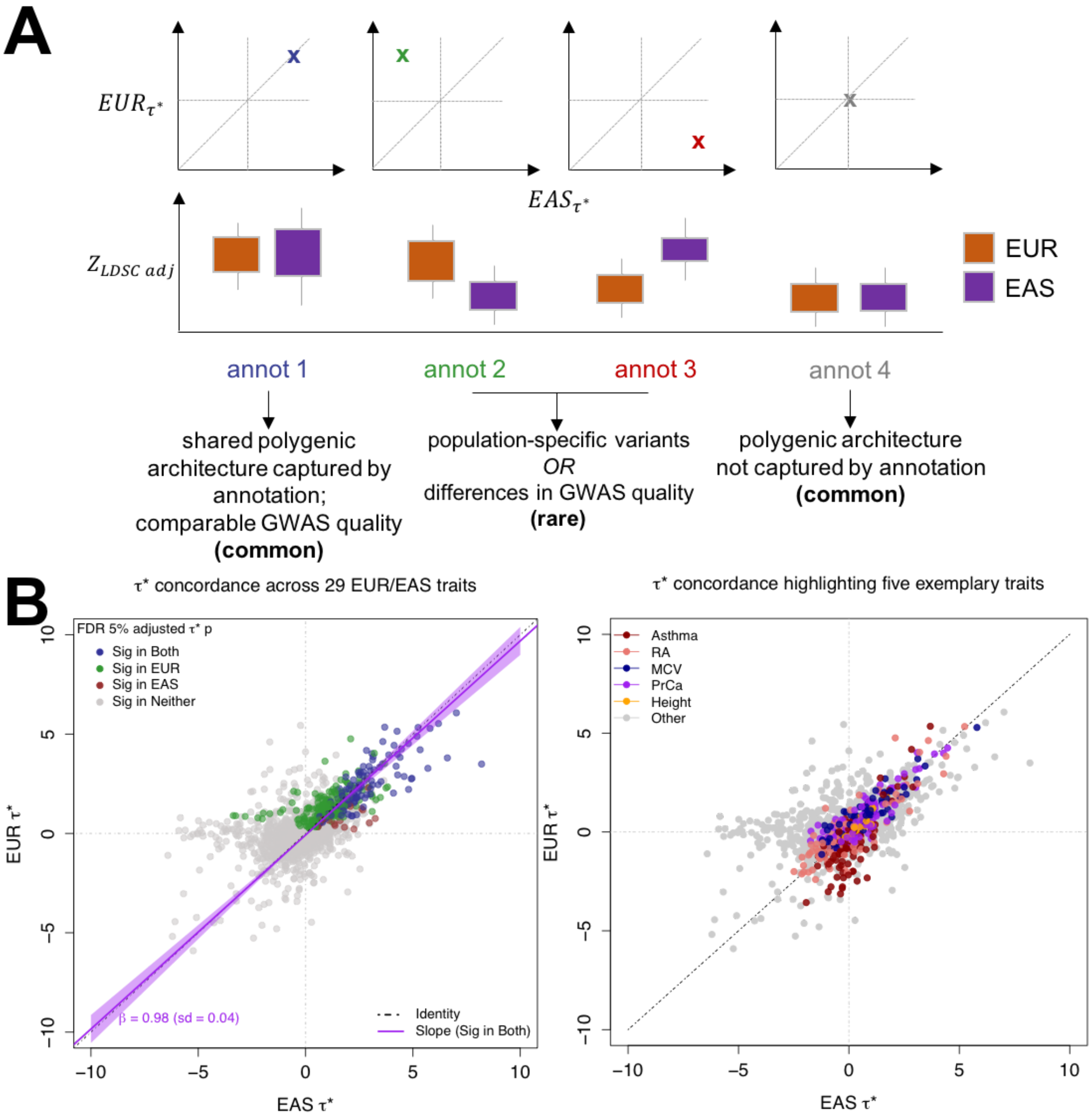


Figure 3-3. Multi-ethnic concordance of regulatory elements defined by IMPACT. A) Illustrative concept of concordance versus discordance of τ * between populations. Concordance implies a similar distribution of causal variants and effects captured by the same annotation. The implications of discordant τ * are not as straightforward. B) Common per-SNP heritability (τ *) estimate for sets of independent IMPACT annotations across 29 traits shared between EUR and EAS. Left: color indicates τ * significance (τ * greater than 0 at 5% FDR) in both populations (blue), significant in only EUR (green), significant in only EAS (red), significant in neither (gray). Line of best fit through annotations significant in both populations (dark purple line, 95% CI in light purple). Black dotted line is the identity line, y = x. Right: color indicates association to one of five exemplary traits.

Assessing variant prioritization with IMPACT toward improving polygenic risk score models

PRS models have great clinical potential: previous studies have shown that individuals with higher PRS have increased risk for disease[18,19,84–86]. In the future, polygenic risk assessment may become as common as screening for known mutations of monogenic disease, especially as it has been shown that individuals with severely high PRS may be at similar risk to disease as are carriers of rare monogenic mutations[86]. However, since PRS heavily rely on GWAS with large sample sizes to accurately estimate effect sizes, there is specific demand for the transferability of PRS from populations with larger GWAS to populations underrepresented by GWAS[18,80,81,83,91,92,96]. As we would like to investigate the ability of IMPACT annotations to improve the trans-ethnic application of PRS, we chose pruning and thresholding (P+T) as our model[18,81]. P+T models, as the name suggests, select an independent subset of all SNPs genome-wide by pruning away SNPs correlated by LD and then further thresholding on GWAS *P* value. We elected to use P+T rather than LDpred[83,96] or AnnoPred[95], which compute a posterior effect size estimate for all SNPs genome-wide based on membership to functional categories. With P+T, we can partition the genome by IMPACT-prioritized and deprioritized SNPs, whereas the assumptions of the LDpred and AnnoPred models do not support the removal of variants, making it difficult to directly assess improvement due to IMPACT prioritization. Moreover, these models have not been explicitly designed or tested for the trans-ethnic application of PRS and thus are beyond the scope of our work. We conventionally define PRS as the product of marginal SNP effect size estimates and imputed allelic dosage (ranging from 0 to 2), summed over M SNPs in the model. Conventional P+T utilizes marginal effect size estimates and therefore is susceptible to selecting a tagging variant over the causal one guided by GWAS *P* values which are inflated by LD. Therefore, we hypothesized that any observed improvement due to incorporation of IMPACT annotations could result from prioritization of variants with higher marginal multi-ethnic effect size correlation (**Figure 3-1D-II**), suggesting these SNPs are less likely to be solely associated by linkage.

Hence, we tested this hypothesis before assessing PRS performance. We selected 21 of 29 summary statistics shared between EUR and EAS with an identified lead IMPACT association in both populations. Then, using EUR lead IMPACT annotations for each trait (**Table B-9**), we partitioned the genome in three ways: (1) the SNPs within the top 5% of the IMPACT annotation, (2) the SNPs within the bottom 95% of the IMPACT annotation, and (3) the set of all SNPs genome-wide (with no IMPACT prioritization). We then performed stringent LD pruning $(r^2$ < 0.1 from EUR individuals of phase 3 of 1000 Genomes[106]), guided by the EUR GWAS *P* value, to acquire sets of independent SNPs in order to compute a EUR-EAS marginal effect size estimate correlation (**Material and Methods**).

For example, in height, EUR-EAS effect size estimates of SNPs in the top 5% partition are 2.1-fold more similar (Pearson *r* = 0.29, **Figure 3-4A**) than those in the bottom 95% partition (*r* = 0.14, **Figure 3-4B**) and 1.6-fold more similar than the set of all SNPs (*r* = 0.18). For each of

17 GWAS *P* value thresholds, the marginal multi-ethnic effect size correlation among the top 5% of IMPACT SNPs tended to be greater than the set of all SNPs genome-wide across 21 traits (all 17 one-tailed paired wilcoxon $P < 6.9e-4$) (**Figure 3-4C-D**). Furthermore, this observation was consistent across individual traits (**Figure B-13**). For comparison, we performed a similar analysis restricted to the five representative traits using alternative functional annotations: lead annotations from 513 cell-type-specifically expressed gene sets (SEG)[25] and 220 cell-type-specific histone mark annotations (CTS)[24] (**Figure B-14**). Marginal effect size correlation with IMPACT was comparable to CTS when comparing the top 5% of SNPs to the set of all SNPs (at each of 17 GWAS *P* value thresholds, one-tailed paired wilcoxon $P > 0.16$, **Figure B-15**). Similarly assessing marginal effect size correlation, IMPACT prioritization was comparable to SEG prioritization (at each of 17 GWAS *P* value thresholds, one-tailed paired wilcoxon $P > 0.06$, **Figure B-15**). Overall, our results suggest that we might anticipate improved trans-ethnic portability of PRS models by prioritizing SNPs in key functional annotations by decreasing the likelihood of selecting SNPs solely associated by linkage.

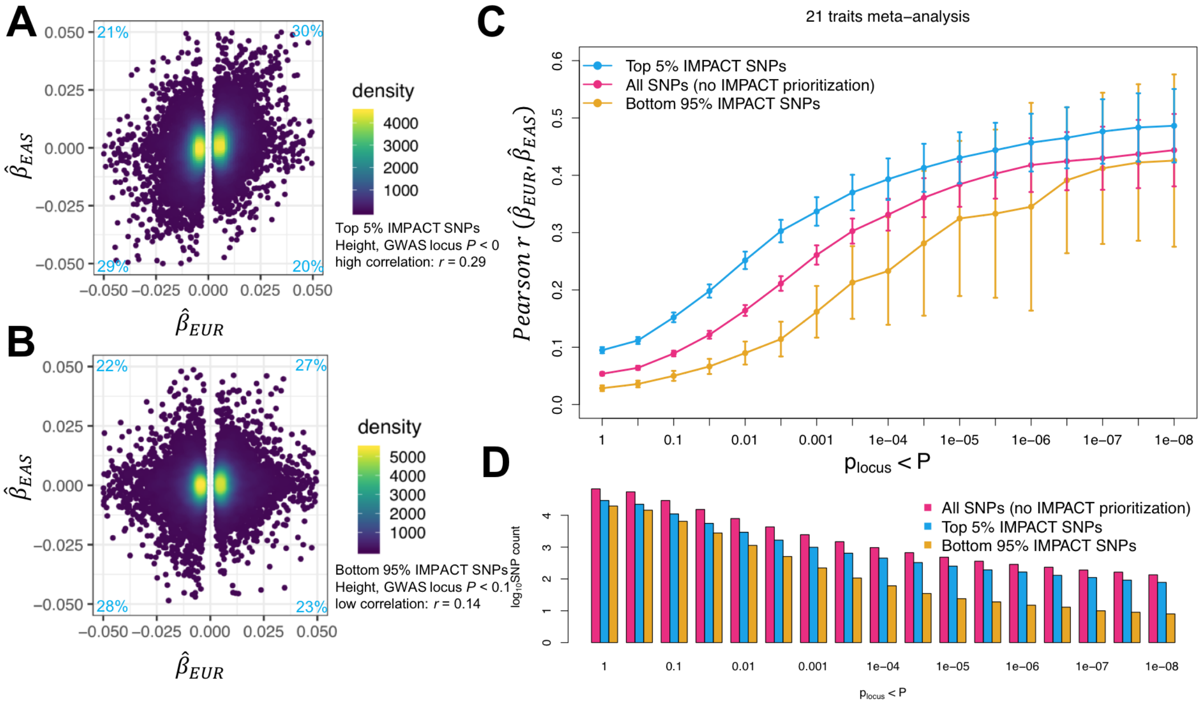


Figure 3-4. Mechanism by which IMPACT prioritization of shared regulatory variants might improve trans-ethnic PRS performance. A) Estimated effect sizes of variants from genome-wide EUR and EAS height summary statistics in the top 5% of the lead IMPACT annotation for EUR height. Proportions of variants in each quadrant indicated in light blue. B) Estimated effect sizes from genome-wide EUR and EAS height summary statistics of variants in the bottom 95% of the same lead IMPACT annotation for height; mutually exclusive with SNPs in A). C) Meta-analysis of multi-ethnic marginal effect size correlations between populations across 21 traits shared between EUR and EAS cohorts over 17 GWAS *P* value thresholds (with reference to the EUR GWAS). Vertical bars indicate the 95% CI around the Pearson *r* estimate. D) Number of SNPs (log10 scale) at each *P* value threshold for each partition of the genome corresponding to C).

While increased concordance of marginal effect size estimates might lead to improved trans-ethnic portability, increased concordance of allelic heterozygosity could also play a role, as allele frequency greatly affects disease predictive power. To this end, we computed the correlation of EUR and EAS heterozygosity (**Material nd Methods**), defined as 2pq, across the same sets of variants and traits considered in **Figure 3-4**. We observed IMPACT-selected variants tended to have lower concordance of heterozygosity than conventional P+T selected variants for each of 17 GWAS *P* value thresholds across 21 traits (all one-tailed paired wilcoxon *P* < 0.05, **Figure B-16, Figure B-17**). This is likely due to an enrichment of common variants among IMPACT-prioritized SNPs and a depletion of rare or low frequency variants (**Figure B-16**). We then considered F_{st}, a measure of the reduction of heterozygosity and an indicator of population divergence, among IMPACT-selected SNPs (**Material and Methods**). Although F_{st} trended higher among IMPACT-selected SNPs than among conventional P+T selected variants across 21 traits at each *P* value threshold (all one-tailed paired wilcoxon *P* < 0.03), the large confidence intervals of the meta-analyzed F_{st} across traits suggest that this trend does not indicate substantial differences (across each of 17 *P* value thresholds, all two-tailed difference of means *P* > 0.98, **Figure B-18, Figure B-19**). These results suggest that neither increased concordance of heterozygosity nor substantial difference in F_{st} is a consequence of IMPACT prioritization.

Models incorporating IMPACT functional annotations improve the trans-ethnic portability of polygenic risk scores

Finally, we addressed our hypothesis that IMPACT annotations improve the trans-ethnic portability of PRS (**Figure 3-1D-III**). For each of the 21 previously analyzed traits, we built a PRS using effect size estimates from EUR summary statistics and applied it to predict phenotypes of EAS individuals from BioBank Japan (BBJ) (**Figure 3-5A**). Here, we compare two PRS models, both blind to any EAS genetic or functional information and removing SNPs with LD r^2 > 0.2, according to European individuals from phase 3 of 1000 Genomes[106]: (i) standard P+T PRS and (ii) functionally-informed P+T PRS using a subset of SNPs prioritized by the lead EUR IMPACT annotation (**Material and Methods**). In functionally-informed PRS models, for each trait separately, we *a priori* selected the subset of top-ranked IMPACT SNPs (top 1%, 5%, 10%, or 50%) which explained the closest to 50% of common SNP heritability (**Material and Methods**). This ensures that functional prioritization captures approximately the majority of trait-relevant genetic variation and the cumulative genetic signal among functionally-prioritized variants was consistent across traits, allowing for varying degrees of polygenicity. For all PRS models, we report results from the most accurate model across nine EUR GWAS *P* value thresholds.

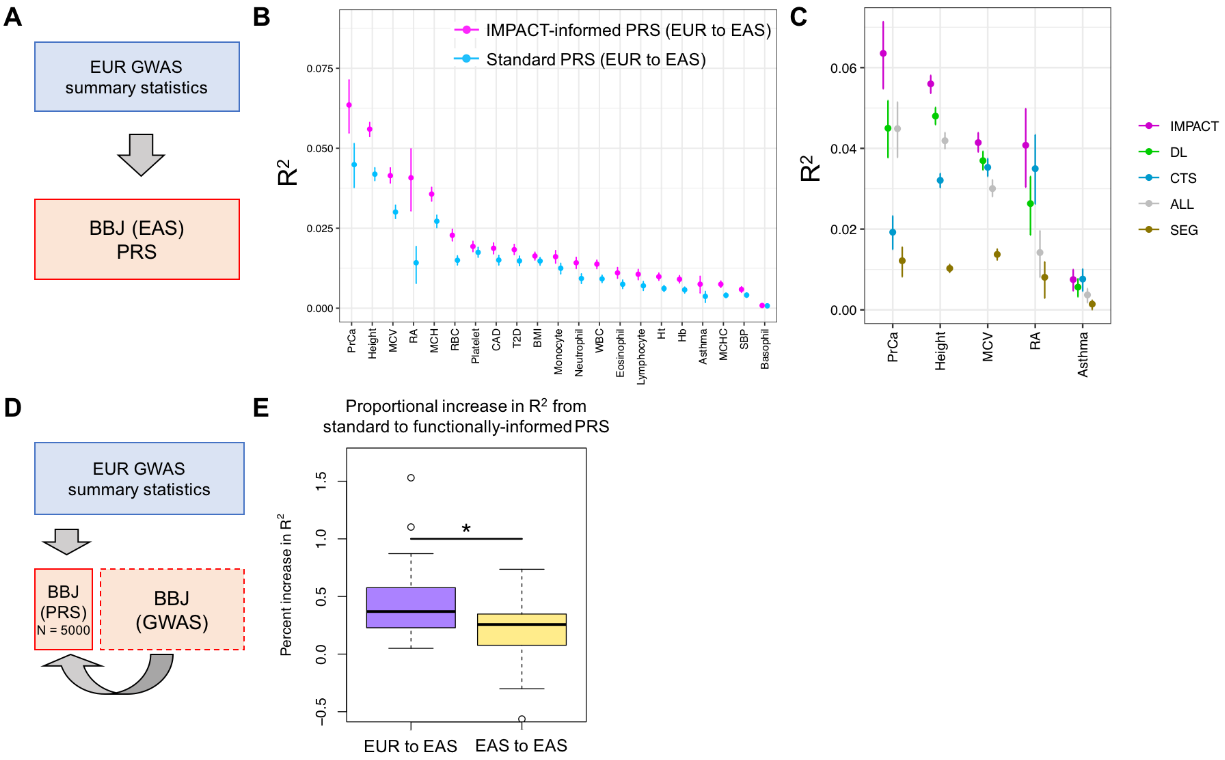


Figure 3-5. Identifying shared regulatory variants with IMPACT annotations to improve the trans-ethnic portability of PRS. A) Study design applying EUR summary statistics-based PRS models to all individuals

in the BBJ cohort. (B) Phenotypic variance (R^2) of BBJ individuals explained by EUR PRS using two methods: functionally-informed PRS with IMPACT (pink) and standard PRS (blue). Error bars indicate 95% CI calculated via 1,000 bootstraps. C) Phenotypic variance (R^2) of BBJ individuals across 5 exemplary traits explained by EUR IMPACT annotations relative to lead deep learning annotations (DL), cell-type-specific histone modification annotations (CTS), and lead cell-type-specifically expressed gene sets (SEG). Error bars indicate 95% CI calculated via 1,000 bootstraps. D) Study design to compare trans-ethnic (EUR to EAS) to within-population (EAS to EAS) improvement afforded by functionally-informed PRS models. For each trait, 5,000 randomly selected individuals from BBJ designated as PRS samples. Remaining BBJ individuals used for GWAS to derive EAS summary statistics-based PRS; no shared individuals between GWAS samples and PRS samples. E) Improvement from standard PRS to functionally-informed PRS compared between trans-ethnic (EUR to EAS) and within-population models (EAS to EAS) using the study design in D). In boxplots, center line indicates the median value; box limits indicate the upper (third) and lower (first) quartiles; the length of whiskers indicate values up to 1.5 times the interquartile range in either direction.

For each of 21 tested traits, we observed that functionally-informed PRS using IMPACT captured more phenotypic variance than standard PRS (49.9% mean relative increase in R^2, **Figure 3-5B**, **Figure B-20**, **Tables B-16-18**). The mean phenotypic variance explained across traits by functionally-informed PRS (R^2 = 2.1%, sem = 0.4%) was greater than by standard PRS (R^2 = 1.5%, sem = 0.3%, one-tailed paired wilcoxon P < 4.8e-7). For 19 of 21 traits, IMPACT-informed PRS significantly outperformed standard PRS (19 one-tailed difference of means P < 0.05); for platelet count P = 0.052 and for basophil count P = 0.40. Using 10,000 bootstraps of the PRS sample cohort, we found that the IMPACT-informed PRS R^2 estimate was consistently greater than the standard PRS estimate for all traits except basophil count (all bootstrap P < 0.004, **Table B-18**). Intriguingly, we found a strong correlation between the

IMPACT-informed PRS R^2 estimate and the EAS heritability captured by the top 5% of SNPs according to the lead EUR IMPACT annotation (Pearson *r* = 0.60, *P* = 0.004, **Table B-19**). While EAS heritability metrics did not influence the choice of lead IMPACT annotation (EUR-based), this result is unsurprising given the strong multi-ethnic regulatory concordance we observed previously (**Figure 3-3C**) in which annotations that capture more heritability in EUR tend to capture more in EAS. Even though IMPACT-informed PRS models include between 7.5% and 79.1% of the total number of SNPs included in standard P+T models, the increased prediction R^2 indicates that prioritization of putatively functional variants over tagging variation compensates for the reduction of included loci. We observed the largest improvement for RA from R^2 = 1.4% (sd = 0.33%) in the standard PRS to R^2 = 4.1% (sd = 0.53%, one-tailed difference of means *P* < 9.8e-6) in the functionally-informed PRS using the B cell TBP IMPACT annotation. For asthma, R^2 = 0.37% (sd = 0.10%) in the standard PRS versus R^2 = 0.75% (sd = 0.14%, *P* < 0.013) in the functionally-informed PRS. For MCV, R^2 = 3.0% (sd = 0.10%) in the standard PRS versus R^2 = 4.1% (sd = 0.12%, *P* < 1.2e-13) in the functionally-informed PRS. For PrCa, R^2 = 4.5% (sd = 0.36%) in the standard PRS versus R^2 = 6.4% (sd = 0.45%, *P* < 6.1e-4) in the functionally-informed PRS. For height, R^2 = 4.2% (sd = 0.10%) in the standard PRS versus R^2 = 5.6% (sd = 0.12%, *P* < 8.7e-20) in the functionally-informed PRS. We observed significantly greater PRS improvement among traits with lower estimated polygenicity (linear regression coefficient = -0.02, *P* < 0.006). As previously stated, more highly polygenic traits may be driven by multiple cell types, of which only one may be captured by the lead IMPACT annotation.

For our five representative traits asthma, RA, MCV, PrCa, and height, we further compared functionally-informed PRS$_{\mathrm{EUR}}$ using IMPACT to models using 123 DeepSEA and

Basenji deep learning annotations[107–110], 220 cell-type-specifically expressed genes (SEG)[25] and 513 cell-type-specific histone modification tracks (CTS)[24] (**Figure 3-5C**, **Table B-11, Table B-20, Material and Methods**). To our knowledge, deep learning annotations have not previously been applied to improving PRS model performance. IMPACT explained greater phenotypic variance on average (mean R^2 = 4.2%, sem = 1.0%) than the top deep learning annotations (3.2%, sem = 0.8%, one-tailed paired wilcoxon *P* = 0.03) and was a significant improvement for four of five traits (four one-tailed difference of means *P* < 0.006), while only trending higher for asthma (*P* = 0.13). IMPACT also explained greater phenotypic variance on average than SEG (0.9%, sem = 0.2%, one-tailed paired wilcoxon *P* = 0.03) and this difference was individually detected for each of five traits (all one-tailed difference of means *P* < 3.4e-6). This trend was not as strong when comparing IMPACT to CTS (R^2 = 2.6%, sem = 0.5%, one-tailed paired wilcoxon *P* = 0.06), although this difference was individually detected for three of five traits (three one-tailed difference of means *P* < 1.1e-4). We performed a similar bootstrap analysis as above, yielding similar results; for only RA and asthma did IMPACT-PRS not produce consistently greater R^2 estimates than CTS-PRS (**Table B-20**).

Functionally-informed PRS might to some extent compensate for population-specific LD differences between populations. Hence, we hypothesized that IMPACT-informed PRS would improve standard PRS moreso in the trans-ethnic prediction framework, in which EUR PRS models predict EAS phenotypes, than in a within-population framework, in which EAS PRS models predict EAS phenotypes. Here, we define within-population PRS as PRSEAS and trans-ethnic PRS as PRSEUR to avoid confusion. In order to directly compare PRS model improvements between PRSEAS and PRSEUR, we evaluated prediction accuracy on the same individuals. Briefly, we partitioned the BBJ cohort to reserve 5,000 individuals for PRS testing, derived GWAS summary statistics from the remaining individuals, and performed P+T PRS modeling and prediction as done above (**Figure 5D, Figures B-21-23, Tables B-21-22, Material and Methods**). For functionally-informed PRSEAS, we selected lead IMPACT annotations from S-LDSC results using GWAS summary statistics, as done above, on the partition of the BBJ cohort excluding the 5,000 PRS test individuals. We defined improvement as the percent increase in R^2 from standard to functionally-informed PRS; therefore, differences in PRS performance due to intrinsic factors, such as GWAS power or genotyping platform, cancel out. In both scenarios, we observed substantial positive improvements: averaged across the 21 traits in the trans-ethnic setting (mean percent increase in R^2 = 47.3%, sem = 8.1%, one-tailed z test P < 2.7e-9) and in the within-population setting (mean percent increase in R^2 = 20.9%, sem = 6.6%, one-tailed z test P < 7.5e-4). Indeed, this revealed a significantly greater improvement in the trans-ethnic application than in the within-population application across the 21 traits (one-tailed paired wilcoxon P < 0.012, **Figure 3-5E**). To ensure that the disease predictive power of our PRS models was not driven by a few loci of large effect, we performed a block jackknife over the genome to establish confidence intervals around the R^2 estimates as well as the relative improvement of IMPACT PRS over standard P+T PRS R^2 estimates (**Material and Methods, Figure B-24**). We observed narrow intervals around the estimates; for functionally-informed PRSEUR and functionally-informed PRSEAS, we observed the average 95% confidence interval around R^2 estimates to be 0.001 and around the relative R^2 improvement to be 0.11 in PRSEUR and 0.07 in PRSEAS. These results suggest that the disease predictive power of IMPACT-informed P+T models are not driven by a few loci of large effect. Moreover, our results for case/control diseases are not affected by estimating marginal effect sizes on the logistic scale, rather than the liability scale[119] (**Material and Methods, Figure B-25, Figure B-26, Appendix B**).

Overall, our results reveal that functional prioritization of SNPs using IMPACT improves both trans-ethnic and within-population PRS models, but is especially advantageous for the trans-ethnic application of PRS. We believe there are at least three important mechanisms at play leading to this improvement. First, restricting P+T PRS to variants that are more likely to be functional increases the likelihood of selecting a causal variant with disease predictive power in the target population. Previous studies support that the identification of causal variants can improve PRS accuracy[81,83,144]. Second, as shown in **Figure 3-3B**, the per-SNP heritability captured by IMPACT annotations tends to be similar in EUR and EAS populations, thereby ensuring that IMPACT-informed SNP prioritization schemes using EUR data are still effective in EAS. Third, as shown in **Figure 3-4C**, SNPs prioritized by IMPACT have more consistent multi-ethnic marginal effect sizes, which means that these SNPs are less likely to be solely associated by linkage and therefore might improve performance. In conclusion, our results nominate the prioritization of SNPs according to functional annotations, especially using IMPACT, as a potential tentative solution for the lack of trans-ethnic portability of PRS models. While individuals of European ancestry dominate current genetic studies, population-nonspecific cell-type-specific IMPACT annotations can help transfer highly powered EUR genetic data to study still underserved populations.

Discussion

In this study, we created a compendium of 707 cell-type-specific regulatory annotations (**Web Resources**) capturing disproportionately large amounts of polygenic heritability in 95 complex traits and diseases in EUR and EAS populations. We then proposed a three-step framework to assess how well prioritization of regulatory variants with functional data can improve multi-ethnic genetic comparisons. First, we showed that heritability-enriched regulatory elements between EUR and EAS populations capture indistinguishable proportions of heritability across 29 complex traits. Second, we showed that functional prioritization of variants selects those with more highly correlated marginal effect sizes between populations, while negligibly affecting the distribution of F_{st} ; this might explain the improvement driven by functional prioritization in P+T PRS models which use marginal effect sizes. Third, we showed that variant prioritization with IMPACT annotations results in consistently improved PRS prediction accuracy, especially for the trans-ethnic application; potentially due to overcoming large population-specific influences such as LD which is an important challenge of multi-population models.

Designing genetic models for each complex trait or disease that capture risk for the full diversity of the human population will be challenging. This necessitates approaches that effectively transfer predictive genetic information from well studied populations to less well studied populations. Without such approaches, the potential clinical benefits of PRS risk to preferentially benefit populations with larger training GWAS datasets, e.g. European populations. As it will ultimately be useful to develop PRS scores that can be applied widely to many populations and admixed individuals[145,146], IMPACT may have the potential to be a tool that can prioritize key variants for this purpose. We argue for the use of biologically diverse IMPACT annotations to capture relevant genetic signal and compensate, to some extent, for differences in LD across populations. To begin to address this, we investigated PRS using EUR summary statistics and genotyping data from five populations (AFR, AMR, EAS, EUR, and SAS) in 1000 Genomes and found that IMPACT-informed PRS moderately reduces the inter-population variation of PRS values compared to standard P+T (one-tailed paired wilcoxon *P* = 0.003, 52.0% reduction in mean F-statistic for EUR PRS (**Figure B-27**) and one-tailed paired wilcoxon *P* = 0.002, 64.6% reduction in mean F-statistic for EAS PRS (**Figure B-28**)), suggesting functional prioritization can stabilize PRS values (**Material and Methods**). However, other challenges such as differences in allele frequencies will need to be addressed in future studies.

Our work and that of others advocate for larger genetic studies in understudied populations[81] and the use of orthogonal LD-independent functional data to improve the disease predictive power of genetic models in such populations, as even increasing GWAS power cannot mitigate the bias introduced by LD. Our study should not in any way be interpreted as a justification for reducing the emphasis on the need for diversity in human genetic studies. A future which offers high powered GWAS in understudied populations will transform the study of trans-ethnic portability from an issue of EUR-biased health disparities to a question of population-specific genetic and environmental effects.

Our work provides insight into the potential clinical implementation of PRS and broader genetic applications that aim to integrate multi-ethnic data. This study suggests that functional data may be leveraged to improve portability of genetic models; however, the issue of portability need not be restricted to two different continental populations as shown in this study, but rather will be relevant to any PRS model in which the target individual is not perfectly matched to the ancestry of the training population. While we did not assess a PRS model using meta-analyzed summary statistics from two or more populations in this study, we believe that this approach could be effective in identifying shared regulatory variants, especially for populations with limited GWAS sample size.

We believe that IMPACT may prioritize phenotype-driving regulatory variation. We have shown IMPACT to be more effective at capturing genetic variation of complex traits than commonly used functional annotations such as experimentally-derived cell-type-specific histone marks, gene sets, and deep learning regulatory annotations. We hypothesize the utility of

IMPACT comes from 1) cell-type-specificity of TF binding models which locate key classes of regulatory elements and 2) the integration of thousands of experimentally-derived annotations, which presumably removes noise and enriches for biological signal present in each individual annotation. Here, we did not demonstrate the potential utility of IMPACT to perform functional fine-mapping to reduce credible sets beyond our previous work[27], due to lack of sufficient gold standards with causal experimental validation and the limitation to genome-wide significant variants. The specific application of IMPACT in multi-ethnic fine-mapping needs to be further investigated.

We must consider several important limitations of our work. First, our functional insights are limited by biases in publicly available TF ChIP-seq data, as IMPACT cannot evaluate TF-cell type pairs for which training data does not exist. These biases include preference toward workhorse cell lines over primary cells or cell types that are rarer or more difficult to assay. Furthermore, these biases include preference toward TFs with evidence of cell type expression and regulation, specific antibodies, and known sequence motifs for compatibility with IMPACT. These biases directly affect our ability to capture trait-relevant biology, leading to systematically better heritability enrichment for autoimmune diseases and hematological traits for which the relevant cell type is easier to assay, e.g. blood, and worse enrichment for brain-related traits for which the relevant tissue is difficult to assay. Future work may be needed to adapt the IMPACT framework to model the epigenetic signatures of functional marks beyond TF binding to capture a broader array of trait-relevant biological processes. In the future, the cell-type-specific functional training data for IMPACT may be replaced by newer experimental strategies to map enhancers. For example, high-throughput CRISPR screens paired with assays for open chromatin could be used to precisely redefine regulatory landscapes. Second, we used multi-ethnic data to argue for the utility of our approach. However, the robustness of multi-ethnic comparisons for a given phenotype rely on properties surrounding the recruitment of individuals or the exact genotyping platform used in various biobanks, which may result in cohort-bias that inflates within-population PRS prediction accuracy. For example, BBJ is a disease ascertainment cohort, in which each individual has any one of 47 common diseases[111,112]; therefore, BBJ control samples are not comparable to healthy controls of UKBB. Other biases may arise from clinical differences in phenotyping. Also, we only considered a single non-EUR population in this study, although the disparity in trans-ethnic portability, and hence resulting benefit from functional annotations, may be greater in other non-EUR populations. Therefore, the results presented here may only be used to interpret the improved portability of genetic data between EUR and EAS populations. Further work is required to assess potential improvements in portability between EUR and other populations.

In conclusion, we demonstrated that IMPACT annotations improve the comparison of genetic data between populations and trans-ethnic portability of PRS models using ancestrally unmatched data. While a long-term goal of the field must be to diversify GWAS and other genetic studies in non-European populations, it is imperative that genetic models be developed that work in multiple populations. Such initiatives will necessitate the use of population-independent functional annotations, such as IMPACT, in order to capture shared biological mechanisms regulated by complex genetic variation.

Acknowledgements

This work is supported in part by funding from the National Institutes of Health (NHGRI T32 HG002295, UH2AR067677, 1U01HG009088, U01 HG009379, and 1R01AR063759) and the Doris Duke Charitable Foundation grant #2013097.

Declaration of Interests

The authors declare no competing financial interests.

Data Availability

We provide IMPACT-707 annotations at https://github.com/immunogenomics/IMPACT/tree/master/IMPACT707

Code Availability

We provide code to recreate our analyses at https://github.com/immunogenomics/IMPACT/tree/master/IMPACT707/AnalysisCode

Web Resources

1. IMPACT Github repository: https://github.com/immunogenomics/IMPACT
2. IMPACT 707 annotations: https://github.com/immunogenomics/IMPACT/tree/master/IMPACT707
3. Analysis code: https://github.com/immunogenomics/IMPACT/tree/master/IMPACT707/AnalysisCode
4. HOMER: http://homer.ucsd.edu/homer/motif/
5. S-LDSC: https://github.com/bulik/ldsc
6. 1000 Genomes: http://www.internationalgenome.org/
7. Cell-type-specifically expressed gene set annotations and LD scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/
8. Cell-type-specific histone modification ChIP-seq datasets: https://data.broadinstitute.org/alkesgroup/LDSCORE/
9. Plink: https://www.cog-genomics.org/plink2
10. Riken website: http://jenger.riken.jp/en/
11. Price Lab: https://data.broadinstitute.org/alkesgroup/sumstats_formatted/
12. Neale Lab: http://www.nealelab.is/uk-biobank
13. GWAS Catalog: https://www.ebi.ac.uk/gwas/
14. Deep Learning: https://data.broadinstitute.org/alkesgroup/LDSCORE/DeepLearning/

Chapter 4

Leveraging single cell epigenomics and genome-wide fine-mapping to study complex trait and disease genetics

The material in this chapter is unpublished and not peer-reviewed.

Abstract

Complex traits and diseases are often driven by key cell-type-specific regulatory mechanisms. However, canonical trait-associated cell types may be composed of cell-states with variable contribution to trait etiology. Moreover, cell-states across canonical cell types may coordinate trait-driving regulatory activities. Single cell technologies are poised to elucidate complex trait genetics by providing an unprecedented resolution of cell-states within a cell type. However, few studies have successfully linked the polygenic effects measured by GWAS to regulatory mechanisms at single cell resolution. Current single cell assays produce data that is extremely sparse, which creates statistical challenges. We propose a unique solution to leverage genome-wide fine-mapping to interpret chromatin accessibility in single cells. To this end, we collected PolyFun fine-mapping results for 5 immune-mediated diseases and 6 blood quantitative traits and 7,811 publicly available Treg, B cell, and monocyte profiles from single cell ATAC-seq (assay for transposase-accessible chromatin + sequencing). Here, we define cell-specific trait scores as the genome-wide average of fine-mapped associations weighted by chromatin accessibility. After identifying heterogeneity of cell-specific trait scores within and across cell types, we sought to explain this heterogeneity by regressing trait scores on cell-specific regulatory program activity scores. We considered regulatory programs defined by immune-related gene sets or IMPACT cell-type-specific regulatory elements. Across 11 traits and 45 regulatory programs, we identified 133 cell-type-specific trait/regulatory program associations and 3 associations in which two different cell types coordinate regulatory activity. Our study reveals prominent factors of trait etiology and provides a biological and regulatory basis for heterogeneity in trait scores within and across canonical cell types.

Introduction

Within the last decade, single cell genomics technologies have rapidly accelerated the field of functional genomics[147,148]. These technologies assay a diverse set of biological phenomena including gene expression, chromatin accessibility, protein expression, protein binding, whole-genome DNA sequencing, DNA methylation, spatial gene expression, and spatial chromatin accessibility[148]. All of these single cell technologies can offer orthogonal understanding of human disease. As approximately 90% of disease-associated variants found by GWAS reside in the noncoding genome[15,16], technologies that measure regulatory activity genome-wide, e.g. not just at genes, might be most helpful in understanding human disease. The assay for transposase-accessible chromatin followed by high-throughput sequencing of single cells (scATAC-seq)[149] identifies genomic regions where DNA is accessible to regulatory factors. Also, as many noncoding variants are thought to have cell-type-specific effects[17,24], scATAC-seq provides unprecedented oppportunities to interrogate the cell-type-specific accessibility and therefore regulatory potential of associated variants at remarkably high resolution.

Previous studies have demonstrated the utility of bulk ATAC-seq[150,151] to explain regulatory effects on gene expression[152] and genetic effects on complex traits and diseases[153]. However, bulk experiments, in which immunophenotypically identical cells are sorted and assayed[150], average together measurements of regulatory activity and possibly obscure underlying cell-states of varying pathogenicity or contribution to trait etiology. More recent studies have demonstrated the utility of scATAC-seq in identifying both *cis*- and *trans*-acting regulators of gene expression in cell-type-specific contexts[154,155]. Studies attempting to leverage single cell chromatin accessibility data to dissect GWAS associations have applied clustering strategies to create pseudo-bulk populations of single cells and then partitioned complex trait and disease heritability according to functional categories of SNPs in cluster-specific regulatory regions[156,157]. However, approaches utilizing cluster-to-cluster heterogeneity do not leverage the power gained by single cell epigenomic experiments to detect cell-to-cell heterogeneity. Fully leveraging the power of cell-specific measurements, a recent study incorporated genetic fine-mapping with scATACseq, revealing the activity of putatively causal variants within specific hematopoietic lineages[158]. However, this study was limited to consideration of genome-wide significant variants and unlike blood quantitative traits, many complex traits and diseases do not have sufficient numbers of genome-wide significant variants to perform meaningful analyses. Ideally, a polygenic model could have been employed to leverage chromatin accessibility across the full spectrum of risk variants genome-wide.

It is now critical to use scATACseq to identify cell-specific asociations with complex traits and diseases under a polygenic inheritance model using a cluster-free approach. However, this requires overcoming the following challenges: the sparse architecture of scATACseq data and the computational inefficiency of testing thousands of cell-specific, genome-wide functional annotations for heritability enrichment across a set of complex traits and diseases. First, the sparse architecture of scATACseq data may violate the assumptions of a polygenic inheritance model, as count data might reflect an oligogenic architecture due to technical drop out. scATACseq data is inherently sparse due to limitations of DNA copy number; in a diploid genome, the Tn5 transposase can only integrate into 0, 1, or at most 2 copies of DNA. Previous studies have developed methods to address sparsity, most of which use feature selection sometimes paired with imputation in order to prioritize relevant genomic regions *a priori* and summarize accessibility over these regions[159,160]. Second, it is computationally prohibitive to test hundreds of functional annotations across a wide variety of complex traits and diseases using strategies to partition genome-wide SNP heritability, such as stratified LD score regression

(S-LDSC). scATACseq datasets typically assay thousands, if not tens of thousands of cells. Using S-LDSC to first compute weighted LD scores for each specific cell and then test for heritability enrichment against a set of GWAS summary statistics would be intractable. For example, consider the computation of weighted LD scores, which takes ~ 2 CPU minutes, compounded by 22 chromosomes, and by 10,000 cells. In fact, many newly published single cell datasets profile hundreds of thousands of cells across clinical cohorts of modest size[161]. Next, consider we'd like to partition heritability for 100 complex traits and diseases and each analysis takes ~ 2 CPU minutes. Although this can be parallelized, 2.4 million CPU hours would be necessary to perform such an analysis.

Here, we focus on immune cell populations and immune-mediated diseases and traits. We collected relevant genome-wide scATACseq profiles and polygenic fine-mapping data in order to define cell-specific trait scores (**Figure 4-1**). Then, in order to identify regulatory processes associated with the variability in cell-specific trait scores, we computed cell-specific regulatory program scores. We define these scores using cell-type-specific functional annotations, including IMPACT tracks and MSigDB gene sets. Our cluster-free approach leverages the full power of single cell measurements and can identify regulatory mechanisms not only specific to subsets of canonical cell types, but also specific to cell-states implicating coordinated regulation in two or more cell types, i.e. a Treg/B cell cell-state with upregulation of interferon signaling. Our approach to integrate single cell chromatin accessibility with GWAS data has the potential to elucidate the biological mechanisms underpinning complex traits as well as nominate novel candidate disease-driving cell states.

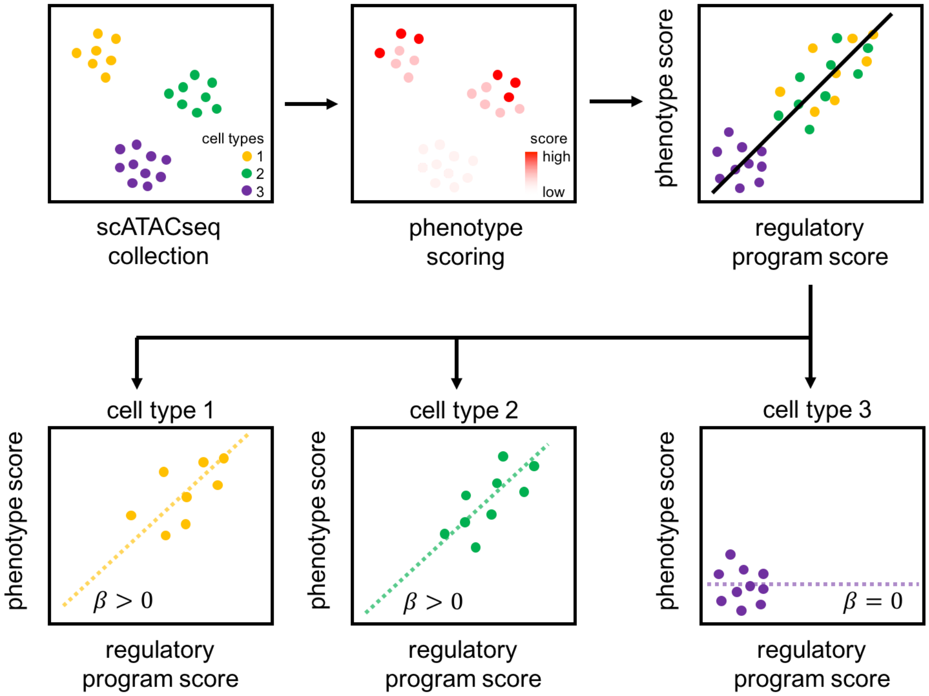


Figure 4-1 legend. Study design schematic. First, we collect publicly available scATACseq data. Second, we define cell-specific trait scores as a genome-wide average of genetic association statistics weighted by chromatin accessibility. Third, we test for associations between cell-specific trait scores and regulatory programs and then characterize the cell-type-specificity of these associations.

Material and Methods

Data

scATACseq data. We collected publicly available scATACseq data consisting of a list of fragments with corresponding cell barcodes and hg19 mapped genomic positions[162]. We collected 3 separate experiments consisting of single cell measurements for 2,661 CD4+ T regulatory cells, 3,155 B cells, and 1,995 monocytes. From this data, we constructed a sparse bin (M = 2,881,044) by cell (N = 7,811) matrix of fragment counts, which we then normalized to represent as a rate of fragments per million (FPM). To obtain 2,881,044 bins, we binned the genome in 1 kb adjacent intervals; intervals do not cross chromosomes. We then filtered out bins with non-zero fragment counts in $\leq$ 10 cells, bins overlapping the ENCODE blacklist (**Web Resources**), and the top 5% of bins with highest representation across the cells to remove invariable regulatory elements such as promoters for housekeeping genes, as previously done[163]. The final bin count in the QC-ed dataset was 622,624 bins.

IMPACT cell-type-specific regulatory element annotation data. We selected 37 independent (r^2 < 0.5) IMPACT annotations, as referenced above, from 140 annotations for which we computed genome-wide base pair resolution scores. These annotations spanned a diverse set of cell types: T cells, B cells, monocytes, PBMCs, myeloid, adipocytes, liver, lung, colon, breast, prostate, stem cell, plasma, myotube, mesendoderm, and ectoderm.

Gene sets. We collected eight separate gene sets representing biological processes related to immune cell functions. First, we defined an interferon (IFN) signature using a panel of 11 genes from a previous study[64]: *HERC5, IFI27, IRF7, ISG15, LY63, MX1, OAS2, OAS3, RSAD2, USP18,* and *GBP5*. Second, we defined an effector T cell cytokine signature using a panel of 8 genes from a previous study[164]: *IL10, TGFB1, TGFB2, TGFB3, IL4, IL5, IL9,* and *IL13.* Third, we downloaded six MSigDB gene sets: IL2-STAT5 (n = 199 genes), Inflammatory response (n = 200 genes), IFN-α (n = 97 genes), IFN-γ (n = 200 genes), TGF-β (n = 54), TNF-α (n = 200 genes).

Statistical Methods

Computing cell-specific regulatory program activity scores defined by IMPACT annotations. Per cell, scATACseq fragment counts follow a negative binomial distribution (**Figure C-1**). To associate IMPACT regulatory programs with single cell accessibility profiles, we build a negative binomial model that regresses cell-specific fragment count data per bin (FPM_i) on *j* independent (r^2 < 0.5) cell-type-specific regulatory element annotations X_j (**Table C-1**), where X_j is a vector of average IMPACT per-nucleotide probabilities per bin for annotation *j* , β_0 is an intercept, μ_i is the mean incidence rate (or risk of additional occurrence) of fragment counts per unit of exposure and t_i is the exposure time[165]:

$$\mu_i \ = \ exp(ln(t_i \ + \ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_j X_j)) .$$

We use a negative binomial regression framework implemented by the *glm.nb* R [v1.0.143] package[58]. We then assess the goodness of fit of the model using a proportion of pseudo-variance explained. In a linear model, the proportion of variance is defined as the squared Pearson correlation between the response variable (y) and the predicted value using the coefficient fits ($\hat{y}$). Here, we computed the squared Pearson correlation between the $log\ (FPM_i \ + \ 1)$ of cell *i* with $X\beta$, the predicted values. We first computed the training-based pseudo-variance explained, which is susceptible to overfitting. To ensure that the regression converged in a timely manner, we downsampled the 622,624 bins that passed filtering (see **Data** below) to 100,000 randomly selected bins genome-wide. We regressed fragment counts on 37 independent IMPACT regulatory element annotations comprising a variety of cell types in a multivariate regression. We next computed the validation set pseudo-variance explained, where training was performed on odd chromosomes to learn β and bins from even chromosomes were used to establish the $log\ (FPM_i + 1)$ and X values in the calculation of the pseudo-variance explained.

Computing cell-specific regulatory program activity scores defined by gene sets. We defined cell-specific regulatory scores according to average chromatin accessibility at sets of genes related to immune function as follows:

$$Regulatory\ score_{cell} = \sum_{g=1}^{G} \left(\sum_{b=1}^{B} (\ fragment_{cell,b}) \ / \ B \ \right) / \ G \ ,$$

where G is the total number of genes in each gene set.

Genome-wide cell-specific SNP-level annotation of fragment counts. As our scATACseq data is represented by a bin by cell matrix of fragment counts, we annotated SNPs genome-wide by identifying the bin in which each SNP resides and then annotating that SNP with the cell-specific fragment count of that single cell. Each bin is 1 kb wide and SNPs within the same bin will be assigned the same fragment count. The result is a matrix of SNPs and cells, where the matrix values are fragment counts.

Computing cell-specific mean chi-squared statistics across genome-wide fine-mapping data. With the S-LDSC framework, we may have defined cell-specific functional annotations as binary indications of chromatin accessibility and then tested these annotations against various complex traits and diseases for heritability enrichments. Given the computationally prohibitive nature of using S-LDSC to test thousands of cell-specific functional annotations, we needed to acquire genetic association data that 1) could be expeditiously analyzed against thousands of cell accessibility profiles and 2) was not susceptible to statistical inflation by LD, which is explicitly regressed out in the S-LDSC framework. We collected 11 previously published genome-wide fine-mapped summary statistics from the PolyFun method[166], which considers all loci genome-wide instead of exclusively genome-wide significant loci as do most fine-mapping studies[16,56]. PolyFun uses the S-LDSC framework to specify prior causal probabilities (proportional to the per-SNP heritability explained by the variant) as the input for conventional fine-mapping tools, such as Susie[167] and FINEMAP[168]. This makes PolyFun rapidly scalable to millions of variants genome-wide. These datasets include variants with a posterior squared effect size > 1e-8, with an absolute value of posterior effect size > 1e-4, that are found in any 95% credible set, or that have a posterior inclusion probability (PIP) > 0.01. For each of 11 datasets and across all included variants, we compute a posterior chi-squared association statistic as the square of the mean posterior effect size estimate divided by its standard error. We define a cell-specific trait score as the genome-wide average across *M* variants of fine-mapped association statistics weighted by cell-specific chromatin accessibility quantified by fragment counts:

$$Mean\ \chi^2_{cell} = \sum_{i=1}^{M} (\chi^2_i * fragment_{cell,i} / \sum_{i=1}^{M} fragment_{cell,i}) / M \ .$$

We establish a null distribution for each trait by shuffling the fragment count matrix 10,000 times.

Linear models to identify regulatory program associations with trait scores. We were interested to know if variability in cell-specific trait scores could be explained by regulatory program activity at the single cell level. We regressed cell-specific trait scores, as computed above, on various cell-specific regulatory program scores. Calculation of IMPACT cell-type-specific regulatory program scores and gene set program scores are described above. We first tested for global association signal, e.g. across all cells, using the following model:

$$y = F + Lab + X ,$$

where y is a vector of cell-specific trait scores, F is a vector of cell-specific total fragment counts over all bins, Lab is a categorical vector of cell type labels: Tregs, B cells, Monocytes, and X is a vector of cell-specific regulatory program scores.

Then, we identified cell-type-specific associations, e.g. either further characterizing global associations or identifying associations that were not detected globally. To do this, for each phenotype and regulatory program pair that we tested, we created three separate cell type models. We elected to use separate cell type models, as opposed to modifying the linear model to include cell type interaction terms. This was because the fragment count variable was modestly correlated with the cell type label, with monocytes having greater fragment counts than Tregs and B cells, which made the designation of reference cell type for the interaction terms bias the estimates.

Results

Cell-specific fragment count-weighted GWAS association statistics

We aimed to leverage the power of single cell resolution chromatin accessibility data to facilitate the identification of noncoding regulatory mechanisms driving complex traits and diseases. Here, we specifically focus on immune cell types and immune-mediated phenotypes. To this end, we acquired polygenic fine-mapping results, e.g. not restricted to genome-wide significant loci, from the PolyFun method for 12 complex traits and diseases from UK BioBank[166]: all autoimmune disease, allergy and eczema, asthma, psoriasis, respiratory ear/nose/throat, eosinophil count, lymphocyte count, mean platelet volume, platelet count, monocyte count, white blood cell count, and body height as a quantitative trait not thought to be associated with these cell types, e.g a negative control. Using posterior effect size estimates and their standard errors, we computed posterior chi-squared association statistics, which are no longer inflated by LD (**Material and Methods**). We also collected 7,811 single cell chromatin accessibility profiles from a previous study[162] composed of 2,661 Tregs, 3,155 B cells, and 1,995 monocytes. For each cell, we defined a cell-specific trait score as the genome-wide average of posterior chi-squared values over approximately 19 million variants fine-mapped with PolyFun, weighted by the proportion of SNP-hitting fragments in that cell and at that variant (**Material and Methods**). We collected cell-specific trait scores for each cell across each of 12 traits. Using trait-specific null distributions of randomized fragment counts over 10,000 trials, we determined the signal-to-noise threshold as cell-specific trait scores that were below the 95th percentile of the null distribution. We estimated such cell-specific trait scores as 0; no more than 0.2% of values for any trait were nullified based on this criterion. To compare cell-specific trait scores between traits that have varying degrees of polygenicity, we normalized these scores by the trait-specific average chi-squared value across the ~19 million variants. We found that meta-analysis of these cell-specific trait scores revealed expected cell-type-specific patterns of biology (**Figure 4-2A**). For example, cell-specific scores for asthma, a T cell-driven disease, were 1.4-fold (one-tailed difference of means *P* < 1.3e-16) and 1.6-fold (one-tailed difference of means *P* < 2.1e-23) larger in Tregs than B cells and monocytes, respectively. Similarly, for allergy and eczema, cell-specific scores among Tregs were 2.1-fold (one-tailed difference of means *P* < 2.4e-34) larger than in B cells and 2.7-fold (one-tailed difference of means *P* < 8.7e-49) larger than in monocytes, respectively. Expectedly, monocyte count scores were 2.3-fold (one-tailed difference of means *P* < 1.9e-50) and 1.4-fold (one-tailed difference of means *P* < 1.2e-14) higher among monocytes than among Tregs or B cells, respectively. Projecting this data into a UMAP using the top 10 PCs revealed heterogeneity among single cells that was obscured by the meta-analysis (**Figure 4-2B-K**). For example, "all autoimmune disease"-high cells seem to implicate subsets of B cells and Tregs (**Figure 4-2C**), platelet count-high cells seem to comprise a specific subset of monocytes (**Figure 4-2E**), and allergy/eczema-high cells account for a fraction of Tregs (**Figure 4-2F**) compared to asthma-high cells which seem to comprise most Tregs (**Figure 4-2G**).

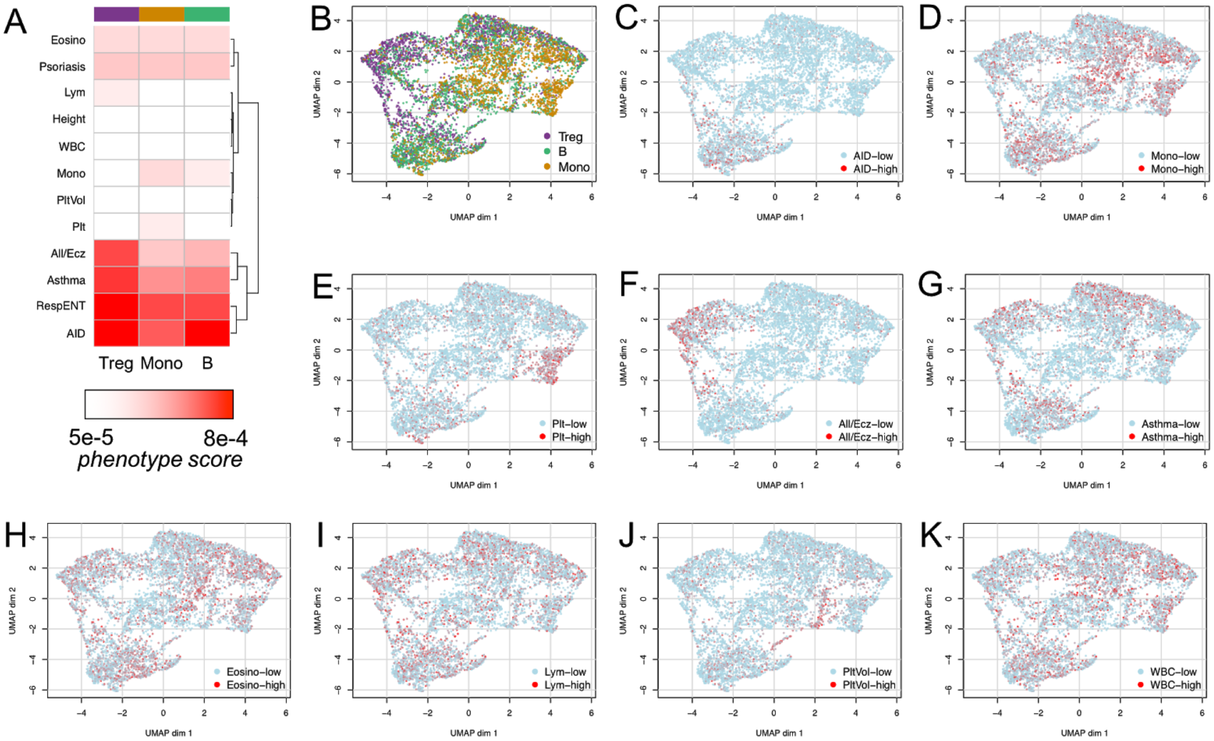


Figure 4-2 legend. Cell-specific trait scores. A) Meta analysis of single cell scores by sorted cell types. B-K) UMAP projection of 10 PCs of cells (N = 7,811) by traits (N = 12) matrix. In B) colors indicate sorted cell type. In C-K) color intensity indicates cell-specific trait scores across nine selected traits, with higher values represented by red and lower values represented by light blue.

Differential enhancer accessibility associated with heterogeneity in cell-specific trait scores

We sought to explain the heterogeneity of cell-specific trait scores observed in **Figure 4-2** by differential accessibility at enhancer elements or gene promoters. To this end, we selected traits with notable heterogeneity in phenotype scores within cell types: “all autoimmune disease” (Tregs and B cells), platelet count (monocytes), and allergy/eczema (Tregs). Then, we dichotomized cells into trait-high and trait-low based on the UMAP and for each 1 kb wide bin genome-wide, we performed a two-tailed wilcoxon test to assess differential accessibility between the two classes of cells. To characterize the differentially accessible bins as putative enhancers, we identified the nearest gene. First, for “all autoimmune disease”, we identified trait-high B cells (n = 893) and trait-low B cells (n = 2,093). We found 4 differentially accessible bins at 5% FDR, all with increased accessibility. Notably, one differentially accessible bin (9.0 log fold change, FDR 5% adjusted *P* < 0.006) was located 186 Mb from the *TYK2* gene, a tyrosine kinase active in interferon signaling pathways and reported to be associated with multiple immunodeficiency diseases include lupus, rheumatoid arthritis, inflammatory bowel disease, type 1 diabetes, Crohn’s disease, multiple sclerosis, psoriasis, primary biliary cirrhosis, and others[127]. We also identified differentially accessible bins between trait-high Tregs (n = 547) and trait-low Tregs (n = 1,995). Consistent with our observations among B cells, the same enhancer element 186 Mb away from *TYK2* was differentially accessible among the Tregs (8.9 log fold change, FDR 5% adjusted *P* < 2.1e-5). Second, for platelet count, we identified trait-high monocytes (n = 354) and trait-low monocytes (n = 1572). We found 53 differentially accessible bins at 5% FDR, all with increased accessibility. Notably, one differentially accessible bin (3.2 log fold change, FDR 5% adjusted *P* = 0.043) was located approximately 6 kb upstream of the *DGKD* gene, a gene that encodes a phosphorylating enzyme that acts on diacylglycerol and has been implicated in a GWAS measuring the proportion of platelets in blood[127]. A second notable bin resides 59 Mb away from *TNFAIP3* (4.5 log fold change, *P* = 0.043) perhaps revealing increased production of the monocyte/macrophage-specific inflammatory cytokine TNF-α. Lastly, a differentially accessible bin located 164 Mb away from *RUNX3* (7.1 log fold change, *P* = 0.049) suggests downregulation of CXCL12, a RUNX3 inhibitor, and its immunosuppressive monocytic functions[169]. Third, for allergy/eczema, we identified trait-high

Tregs (n = 938) and trait-low Tregs (n = 1,604). We found 2 differentially accessible bins at 5% FDR, both with increased accessibility. The first bin (21.1 log fold change, $P < 3.6e\text{-}7$) resides approximately 8 Mb away from the *GAL3ST2* gene, a galactose sulfotransferase, which has been observed to be associated with allergy and asthma in multiple GWAS studies, including childhood onset and adult onset asthma[127]. The second differentially accessible bin (22.8 log fold change, $P < 2.6e\text{-}8$) resides 133 Mb away from the *SMARCE1* gene, which has been implicated by separate GWAS for eczema and asthma[127] and is involved with chromatin remodeling permitting expression of generally repressed genes.

Identifying regulatory programs associated with increased cell-specific trait scores

The identification of differentially accessible bins above has implicated enhancer elements involved in known trait-relevant biological processes. However, since this analysis tested differences at each individual bin genome-wide, we are powered to only detect large effects which are scarce in scATACseq data due to sparsity and relatively low signal to noise ratio. Therefore, we sought to leverage the cumulative changes in accessibility between cell states across many gene loci implicated by various regulatory programs. Thus, we might be able to associate changes in cell-specific trait scores with the activity level of these regulatory programs.

To this end, we defined continuous-valued cell-specific scores representing the activity level of 45 regulatory programs (**Material and Methods**). Of these 45, we investigated 37 independent regulatory programs defined by *in silico* cell-type-specific IMPACT regulatory element annotations across a variety of cell types. To compute cell-specific IMPACT scores, for each of 7,811 single cell chromatin accessibility profiles, we regressed fragment counts on 37 independent IMPACT annotations averaged over genomic bins corresponding to the scATACseq data in a negative binomial regression model (**Material and Methods**). For each cell and for each of 37 IMPACT annotations, we obtained a set of regression coefficients and their standard errors. However, the IMPACT regression models explained a small amount of pseudo-variance (**Material and Methods**) in the fragment count data (within-training: mean = 0.0002, se = 2.0e-5, validation (odd vs even chromosomes): mean = 0.01, se = 5.1e-4). Next, for two regulatory programs, we defined an interferon gene signature [64] and an immune-related cytokine gene signature[164]. Lastly, we downloaded six hallmark gene sets (**Materials and Methods**) related to immune cell regulation from the MSigDB database (**Web Resources**): IL2-STAT5, IFN-α, IFN-γ, TGF-β, TNF-α, and interferon response. We define cell-specific gene set scores as the average fragment count across bins overlapping the corresponding genes.

We then used a linear regression model to assess the association of cell-specific trait scores with each of these regulatory programs, while accounting for total fragment count per cell and cell type as covariates (**Material and Methods**). We then created cell-type-specific models to further characterize the cell-type-specificity of the associated regulatory programs (**Material and Methods**). In total, we identified 76 trait-increasing associations in the cell-type-nonspecific models and 133 trait-increasing associations in the cell-type-specific models with anova $P < 0.05$ at 5% FDR (test model compared to null model). Consistent with what we observed in the bin-based differential accessibility analysis, “all autoimmune disease” scores are significantly associated with greater accessibility at interferon response genes (β = 0.02 (se = 0.009), 5% FDR adjusted P = 0.037), however this association was not detected in the Treg- and B cell-specific analyses. Also consistent with our previous analysis, platelet count scores are significantly associated with greater accessibility at genes involved in the TNF-α pathway (β = 0.04 (se = 0.009), 5% FDR adjusted P = 6.2e-4). Moreover, this TNF-α association was found to be specific to monocytes in the cell-type-specific analysis (β = 0.07 (se = 0.02), 5% FDR adjusted *P* = 6.2e-3), e.g. no detectable association in Tregs or B cells.

Next, we report the top associated regulatory program for each trait (**Table 4-1**). Intriguingly, TNF-α specifically in monocytes was the top associated regulatory program for nine of eleven considered traits. This suggests that monocytes with higher phenotype scores have increased enhancer activity related to the TNF-α regulatory program. This is reasonable as monocytes serve as the primary producers of TNF-α in humans[170]; thus higher TNF-α scores might indicate stronger monocytic identity and function.

Trait	Top regulatory program	beta	SE	FDR-adjusted P <
All AID	TNF-α (Mono)	0.05	0.01	0.006
Eosino count	TNF-α (Mono)	0.08	0.02	0.0001
Lym count	TNF-α (Mono)	0.05	0.02	0.01
Plt Volume	TNF-α (Mono)	0.10	0.02	4.2e-5
Monocyte count	TNF-α (Mono)	0.08	0.02	0.003
Plt count	IFN-α (Mono)	0.07	0.02	0.01
WBC	TNF-α (Mono)	0.08	0.02	8.2e-5
Allergy/eczema	TNF-α (Mono)	0.04	0.02	0.03
Asthma	IFN-γ (Mono)	0.07	0.03	0.04
Psoriasis	TNF-α (Mono)	0.05	0.01	0.002
Resp ENT	TNF-α (Mono)	0.06	0.02	0.005

Table 4-1 legend. Top regulatory program association of gene sets and IMPACT annotations for each of eleven traits from linear regression model. Program further described by the cell-type-specific model in which it was identified, e.g. monocytes. Top association chosen by largest β whose FDR 5% adjusted $P < 0.05$, where the P value is computed using an anova between the null and test models.

To summarize our findings, we separately visualized gene set regulatory program results (**Figure 4-3A**) and IMPACT regulatory annotation results (**Figure 4-3B**). The β attributable to the regulatory programs defined by gene sets (mean = 0.04, sd = 0.002) were 12.8-fold greater on average than those attributable to IMPACT regulatory annotations (mean = 0.003, sd = 0.0001). Generally, we found that associations with TNF-α, IFN-α, IFN-γ, and TGF-β gene sets were often detected in monocyte-specific models and that these associations were among the strongest in magnitude. In fact, about half of these associations were not detected in the cell-type-nonspecific models. These results suggest cell-specific trait score heterogeneity in monocytes is proportional to enhancer accessibility at TNF-α, IFN, and TGF-β genes across a wide range of blood traits and immune diseases. Next, associations with IL2-STAT5 were mostly detected in Treg-specific and B cell-specific models and a fraction of the time only detected in cell-type-nonspecific models. This suggests common regulatory activity within Tregs and B cells in which increased enhancer accessibility at genes related to the IL2-STAT5 pathway explains some of the etiology of both blood traits and immune diseases alike. Next, the smaller gene panels, IFN and T cell cytokines, from previous studies resulted in weaker associations. However, first, the only cell-type-specific associations with the IFN panel implicate B cells, whereas B cells were not implicated by any other IFN-related pathway, e.g. IFN-α, IFN-γ, or interferon response (except in the case of white blood cell count). Second, the cytokine panel, although defined based on effector T cells, was subtly associated in monocyte-specific models with lymphocyte count and “all autoimmune disease”.

Intriguingly, we found that for three trait-program pairs, our strategy detected a global association which is also individually detected in two of three cell types (**Figure 4-4**). Specifically, for platelet count and white blood cell count, we observed an association with IL2-STAT5 genes, an association that was separately detected within Tregs and B cells, but not monocytes. This result is not surprising given the Treg and B cell specificity of other IL2-STAT5 associations observed in **Figure 4-3A**. We also found that white blood cell count scores were associated with increased accessibility at interferon response genes, but this association was separately detected within B cells and monocytes, but not Tregs. This is an interesting result as interferon response was rarely observed to be associated in B cell-specific models, with the exception of WBC, allergy/eczema, and weakly asthma and respiratory ENT.

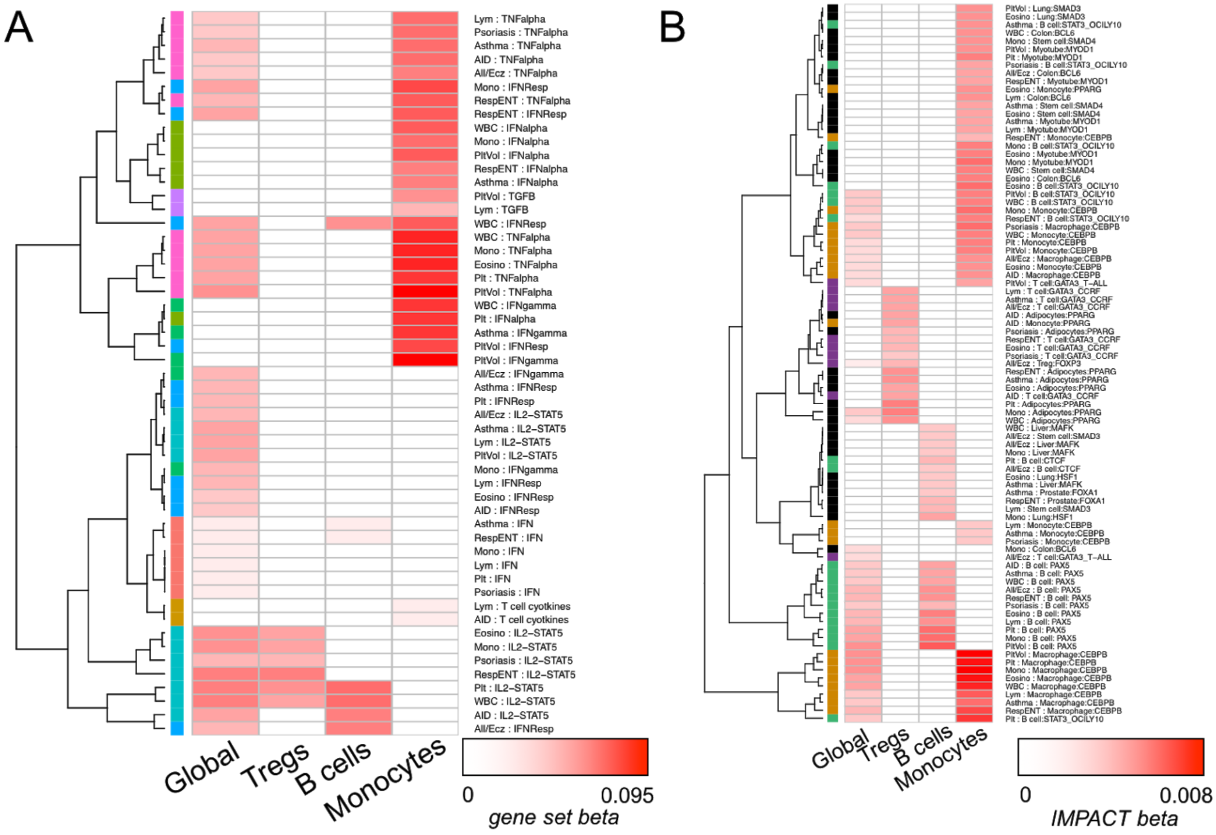


Figure 4-3 legend. A,B) One-dimensional hierarchical clustering on matrix of association β (with linear model 5% FDR-adjusted $P < 0.05$ in anova between null and test models). If $P > 0.05$, β represented as 0. Columns indicate the model regime in which the association β were computed, global indicates the cell-type-nonspecific model (all cells). A) Rows indicate the trait and gene sets program pairs. Legend colors indicate the program. B) Rows indicate the trait and IMPACT regulatory annotation pairs. Legend colors indicate the cell type of the IMPACT annotation.

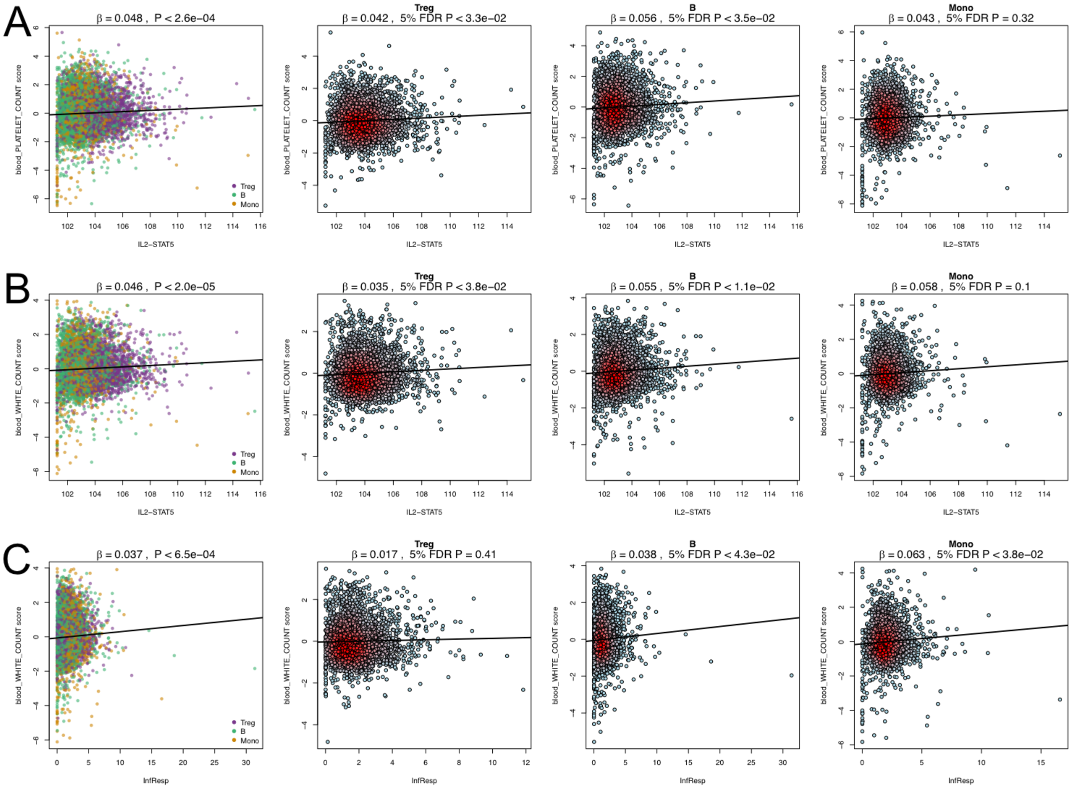


Figure 4-4 legend. Three instances of detectable global trait/program associations in which the same regulatory association is detectable with two of three cell-type-specific models. The IL2-STAT5 program is associated with platelet count in A) and white blood cell count in B) and this effect is shared between Tregs and B cells, but not monocytes. In C) Interferon response is associated with white blood cell count and this effect is shared between B cells and monocytes, but not Tregs.

As for IMPACT cell-type-specific annotation associations (**Figure 4-3B**), we unsurprisingly observed cell-type-specific behavior. Namely, we rarely observed global associations that were not specific to one particular cell type. Also, we did not observe any associations that were shared across cell types. This is expected as IMPACT regulatory annotations implicate cell-type-specific biology. We found that monocyte and macrophage

IMPACT annotations were strongly associated with blood traits and immune diseases alike in monocyte-specific models. Similarly, Treg annotations and B cell annotations were strongly associated in Treg-specific and B cell-specific models, respectively. Unexpectedly, adipocyte IMPACT annotations tended to associate with traits in Treg-specific models, while liver, lung, prostate and stem cell annotations tended to associate with traits in B cell-specific models, and myotube and colon annotations tended to associated with traits in monocyte-specific models. Generally among IMPACT annotations associated in Treg-specific models, we observed greater specificity for immune diseases than blood traits; this was not the case with gene set programs. On the other hand, IMPACT annotations associated in B cell- and monocyte-specific contexts showed no preference for immune diseases or blood traits.

Although the top associations reported in **Table 1** largely implicated monocyte regulation, the heterogeneity of single cell scores in **Figure 4-2** seem to implicate other cell types depending on the phenotype. For example, cells with high “all autoimmune disease” scores are more likely to be Tregs and B cells than monocytes. Therefore, rather than identifying that the TNF-alpha program in monocytes is associated with cells with higher “all autoimmune disease” scores, we would like to identify the top associated program in Tregs or B cells, e.g. the relevant cell types. Therefore, we next identified the top association for the seven traits in which monocytes do not explain the majority of variance of the cell-specific trait scores (**Table 4-2**). These results suggest strong regulatory differences between these immune-mediated diseases. For example, allergy and eczema scores are driven most strongly by Tregs and are concordant with the IMPACT T cell (GATA3) annotation, as GATA3 activity is known to be important for proper Treg function and Tregs have a recognized role in allergy and eczema[171]. On the other hand, asthma scores which also implicate Tregs are most concordant with the adipocyte (PPARG) IMPACT annotation, consistent with airway inflammation of adipose tissue contributing to asthma severity[172]. Lastly, for psoriasis and respiratory ENT, for both of which Tregs play a considerable role, the increased accessibility at IL2-STAT5 genes is most concordant with higher cell-specific scores.

Trait	Top regulatory program	beta	SE	FDR-adjusted P <
All AID	Tregs: T cell (GATA3) B cells: IL2-STAT5	0.003; 0.05	0.0008; 0.02	0.002; 0.04
Lym count	Tregs: T cell (GATA3) B cell: B cell (PAX5)	0.003; 0.004	0.0007; 0.0003	0.003; 2.2e-22
Plt Volume	B cells: PAX5	0.005	0.0004	8.9e-29
Allergy/eczema	Tregs: T cell (GATA3)	0.003	0.001	0.03
Asthma	Tregs: adipocytes (PPARG)	0.003	0.001	0.008
Psoriasis	Tregs: IL2-STAT5 B cells: PAX5	0.03; 0.003	0.01; 0.0003	0.04; 3.3e-16
Resp ENT	Tregs: IL2-STAT5	0.05	0.01	0.004

Table 4-2 legend. Seven traits for which trait-high cells are predominantly represented by Tregs or B cells. Column 2 reports the β from the top association in either Tregs or B cells as indicated. Two sets of metrics reported for traits whose high scoring cells are represented by both cell types.

Discussion

In this study, we leveraged the power of single cell chromatin accessibility profiles to identify cellular heterogeneity in the etiology of polygenic blood traits and immune-mediated diseases. Specifically, we identified and characterized regulatory heterogeneity in canonical cell types (Tregs, B cells, and monocytes), revealing cell-states in which upregulated gene regulatory programs are associated with higher cell-specific trait scores. These cell-specific trait scores represent single cell-level chromatin accessibility at fine-mapped variants and we define them as averaged genome-wide chi-squared statistics weighted by fragment counts. The cell-states we observed are not strictly subpopulations of Tregs, B cells, and monocytes; for example, we identified three instances of shared regulatory profiles across cell types. First, increased single cell chromatin accessibility at IL2-STAT5 pathway genes were associated with increased platelet count scores in both Tregs and B cells, revealing a cell-state that bridges Tregs and B cells. Second and similarly, increased white blood cell count scores were associated with increased IL2-STAT5 scores in both Tregs and B cells, revealing the shared role of Tregs and B cells alike in regulating platelet count and white blood cell count via the IL2-STAT5 pathway. Third, increased white blood cell count scores were associated with increased accessibility at genes related to interferon response, revealing a cell-state bridging B cells and monocytes.

There are some important limitations of this work to consider. First, our analysis of various cell types and complex traits and diseases relied on the availability of relevant data. Due to an abundance of regulatory annotations and relevant blood traits and immune-mediated disease fine-mapping data, we chose to focus this study on several immune cell types: Tregs, B cells and monocytes. Moreover, there may be other important blood or immune cell types to consider in terms of the polygenic traits and diseases considered in this study. Therefore, our trait-specific conclusions may not implicate the most relevant associated regulatory program or cell type, only the most relevant program among the three cell types we studied. Second, our study relies on publicly available scATACseq from a single study and we have not assessed if our results are robust to using other datasets of the same cell type and experimental protocol. Third, we elected to represent scATACseq data using genome-wide adjacent bins rather than calling peaks and counting fragments in peaks for each cell. A desirable feature of peak calling is the establishment of a baseline signal level, and thus our usage of all fragment counts likely includes some degree of technical artifact. However, peak calling also has the undesirable outcome of missing peaks relevant to rare cell types or cell states. However, the sizes of subsets of trait-high or trait-low cells dichotomized in our study (hundreds of cells) might not be small enough to justify this power concern. Fourth, we downloaded only a small number of gene sets from MSigDB and are thus unlikely to have accounted for all relevant regulatory programs that may be regulating the analyzed complex traits and diseases in the analyzed cell types. This calls for a more comprehensive analysis with a greater number of considered gene sets, however it is uncertain as to how this would affect the multiple testing burden as association signals are modest.

Altogether, this work provides a new framework for integrating polygenic fine-mapping, single cell epigenetic assays, and functional annotation data. This multi-modal integration has the potential to identify novel important cell-states and regulatory mechanisms in human complex traits and diseases.

Web Resources

1. ENCODE blacklist: https://personal.broadinstitute.org/anshul/projects/encode/rawdata/blacklists/
2. MSigDb: https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp

Chapter 5

Discussion

With this thesis, I hope to have demonstrated the versatility and broad genomic applicability of IMPACT *in silico* functional annotations. Moreover, I hope to have convinced the reader of the importance of designing high quality functional annotations, underscoring the importance of high quality data generation. I've shown that predictive modeling to create *in silico* functional annotations complement the analysis of many other biological data types. First, improved functional annotation provides a biologically-relevant and orthogonal basis for the analysis of genetic association data, such that we might overcome inherent statistical confounding, e.g. from LD and other population-specific genomic structure. Second, *in silico* functional annotations provide a way to leverage thousands of generated experimental datasets in a way that smooths out the noise and variability of any single dataset by identifying consistent patterns relevant to a biological process, e.g. *in vivo* TF binding. Third, this type of predictive modeling allows us to make inferences that we could otherwise not make using publicly available experimentally-generated datasets, e.g. where other cell type regulation or TF binding might be occurring despite only measuring occupancy of one TF. Fourth, cell-type-specific functional annotation can help generate hypotheses regarding regulatory mechanisms, assisting experimental design for follow-up analysis.

My dissertation work does not fully encompass all scientific avenues of application of these annotations. To encourage continued use of these annotations beyond my thesis work, in this chapter, I will briefly discuss unexplored applications of IMPACT functional annotations.

Functional prioritization of variants to be tested in genome-editing experiments

Recently, advances in genome-editing, including CRISPR, provide the opportunity to test the effects of targeted, allele-specific changes on gene expression, enhancer activity, cell fate, and more. However, two important experimental factors often limit the tractability of these studies. First, often the yield of cells with the intended edit is exceedingly low. Second, it is often unclear which base-pair edit will result in a measurable change, necessitating investigation of a large target region, such as an enhancer. Therefore, hypothesis-driven approaches can reduce the experimental burden of genome-editing. For example, prioritizing regions or variants with high IMPACT regulatory probability in the relevant cell type may expedite the identification of functional elements.

Improved power to detect trans-eQTLs

Trans-eQTLs are variants that modulate gene expression from a distance. For example, a trans-eQTL might reside on one chromosome and the regulated eGene on another. Such interactions are possible due to several biological phenomena. First, long-range chromatin architecture can place a promoter in close 3D proximity to a linearly distal enhancer. Second, such interactions may be indirect: for example, a variant might modulate the expression of a nearby gene and the protein product of that gene might influence the expression of a distal gene. The identification of trans-eQTLs is a statistically challenging problem for several reasons. First, trans-eQTLs often have vastly smaller effect sizes than cis-eQTLs. This limits the power with which we can confidently identify a relation between a trans-eQTL and its eGene. Second, trans-eQTLs tend to be more cell-type-specific than cis-eQTLs; therefore, their discovery highly depends on assaying the relevant cell type. As I've shown in **Chapter 2**: *Improved enrichment of gene expression causal variation*, we found that variants with higher IMPACT cell-type-specific regulatory element probabilities are enriched for cis-eQTL genetic variation. We hypothesized that in an eQTL discovery experiment, using IMPACT probabilities as functional priors might increase power to detect cis-eQTLs, by reducing the number of tested SNP-gene pairs thereby reducing the multiple testing burden. While SNP-gene pairs within some proximal gene window are tested for cis-eQTLs, the multiple testing burden is largest for detection of trans-eQTLs as the remainder of the genome is considered. Therefore, the power to detect trans-eQTLs specifically stands to benefit from cell-type-specific functional priors, such as from IMPACT annotations.

Discovery of unknown cell-type-regulating transcription factors

The epigenetic signature that IMPACT learns might help to identify TFs that are also important regulators of the target cell type. As stated in **Chapter 2**: *Discussion*, TF ChIP-seq datasets are limited to TFs that have been known to regulate particular cell types by prior functional knowledge. Moreover, TFs for which specific antibodies do not exist cannot be assayed. Therefore, in a given cell type, there are potentially many TFs of unrecognized regulatory activity. Using the epigenetic signature learned by IMPACT to identify cell-type-specific regulatory elements, one could then perform motif enrichment of known TFs to identify important regulators or de novo motif discovery to identify important regulatory sequences.

Appendix A

Supplementary Information for Chapter 2

Supplementary Tables

Supplementary Tables A-1, A-2, A-6 and A-7 can be found in the online supplement of Amariuta et al, *AJHG* 2019 and due to their size are not included below.

Table A-3. DNase-seq and or expression data used for benchmarking.

Benchmarking datasets for Mocap	**DNase-seq bam file ENCODE accession**	**Expression ENCODE accession**
CD4+ Th1	ENCSR000EQE	NA
CD4+ Th2	ENCSR000EQG	NA
CD4+ Th17	ENCSR000EQF	NA
CD4+ Treg	ENCSR000EQK	NA
Pancreas (PANC-1)	ENCSR000EPT	NA
Fetal Brain	ENCSR475VQD	NA
Lymphocyte (GM12878)	ENCSR000EMT	NA
Liver (HepG2)	ENCSR000EJV	NA
Benchmarking datasets for Virtual ChIP-seq	**DNase-seq narrow peak file ENCODE accession**	**Expression ENCODE accession**
sigmoid colon	ENCFF862GVN	ENCFF732ORT
fibroblast	ENCFF837VPM	ENCFF777IZN
heart	ENCFF327HDP	ENCFF199GQY
liver	ENCFF286LYP	ENCFF804QWF
Pancreas (PANC-1)	ENCSR000EPT	ENCFF554RBN
stomach	ENCFF151DGI	ENCFF026JIM

Table A-4. Summary and accession information for transcription factor ChIP-seq data.

TF	Total nucleotides covered	Total peaks	Mean nucleotides per peak	NCBI GEO / ENCODE Accession
T-BET	57,708,437	51069	1821	GSM2176974, GSM2176976, GSM1527682
GATA3	5,861,945	23,742	246	GSM1859075
STAT3	3,402,600	5,681	622	GSM2545819
FOXP3	1,354,127	9,847	157	GSM1056936, GSM1056937
STAT5	293,719	2,281	129	GSM1056923, GSM1056922
IRF5	5,020,855	7,640	657	GSE38567
IRF1	2,332,588	7,013	333	GSM1057026
CEBPB	2,177,346	13,694	159	GSM785496
PAX5	5,426,269	15,927	341	GSM1086293, GSM1086294, GSM1086295, GSM1086296
REST	6,309,445	16,404	384	GSM803335
RXRA	8,478,764	13,186	643	GSM1010767
HNF4A	12,395,645	50,421	380	GSM803460
TCF7L2	13,397,935	27694	484	GSM816438
RNA PolII in lymphocytes	38,815,231	71,879	540	GSM1527695, GSM1527697, GSM486494
RNA PolII in sigmoid colon	9,825,888	21,186	464	ENCFF207KQM, ENCFF307VSP, ENCFF534MPD, ENCFF844HAN
RNA PolII in fibroblasts	41,002,445	24,897	1646	GSM1405132
RNA PolII in left ventricle heart	8,375,044	19,704	425	ENCFF596FFJ, ENCFF813MTO
RNA PolII in liver	32,622,340	40,330	809	GSM1010821
RNA PolII in stomach	3,545,545	9,696	366	ENCFF957WEQ, ENCFF168HOV, ENCFF849CHG, ENCFF874HRJ
RNA PolII in pancreas	25,020,627	36,746	681	ENCFF386FNQ, ENCFF628UBC, ENCFF665RHY, ENCFF760CPU
RNA PolII in CD4+ T cells	14,988,606	23,698	632	GSM1527694, GSM1527695

Table A-5. Statistics regarding IMPACT annotations and their relationship to RA polygenic heritability, including annotation size, enrichment, and per-annotation standardized effect size (τ *).

IMPACT Model	**Average value of annotation genome-wide (same for EUR and EAS)**	**EUR enrichment [95%CI], p-value**	**EAS enrichment [95%CI], p-value**	**EUR τ^* [95%CI], p-value**
T-BET (Th1)	0.03 (0.13)	14.6 [8.7 - 20.4], 8.95E-08	17.9 [8.5 - 27.3], 1.57E-05	2.5 [1.4 - 3.7]
GATA3 (Th2)	0.03 (0.09)	12.8 [7.7 - 17.8], 9.11E-08	11.6 [6.2 - 17.0], 1.22E-06	3.8 [2.2 - 5.3]
STAT3 (Th17)	0.01 (0.07)	25.0 [14.0 - 36.0], 1.18E-06	31.2 [12.1 - 50.3], 1.90E-04	3.8 [1.9 - 5.7]
FOXP3 (Treg)	0.02 (0.09)	22.9 [13.6 - 32.3], 5.24E-08	24.4 [10.9 - 37.8], 2.55E-05	4.1 [2.4 - 5.7]
IMPACT Model	**EUR- τ^* [95%CI], p-value**	**EUR- top 5% prop RA h2 (se)**	**EUR- top 5% enrichment p**	**EAS- top 5% prop RA h2 (se)**
T-BET (Th1)	2.6 [1.4 - 3.9]	0.62 (0.12)	1.40E-08	0.79 (0.20)
GATA3 (Th2)	2.6 [1.1 - 4.1]	0.77 (0.15)	4.20E-09	0.76 (0.18)
STAT3 (Th17)	4.6 [2.1 - 7.0]	0.75 (0.15)	6.40E-08	0.87 (0.22)
FOXP3 (Treg)	3.7 [1.8 - 5.6]	0.84 (0.17)	7.20E-08	0.89 (0.22)

Supplementary Figures

Figure A-1. IMPACT parameter selection: lasso (L1) and ridge (L2) mix term and feature distance parameter.

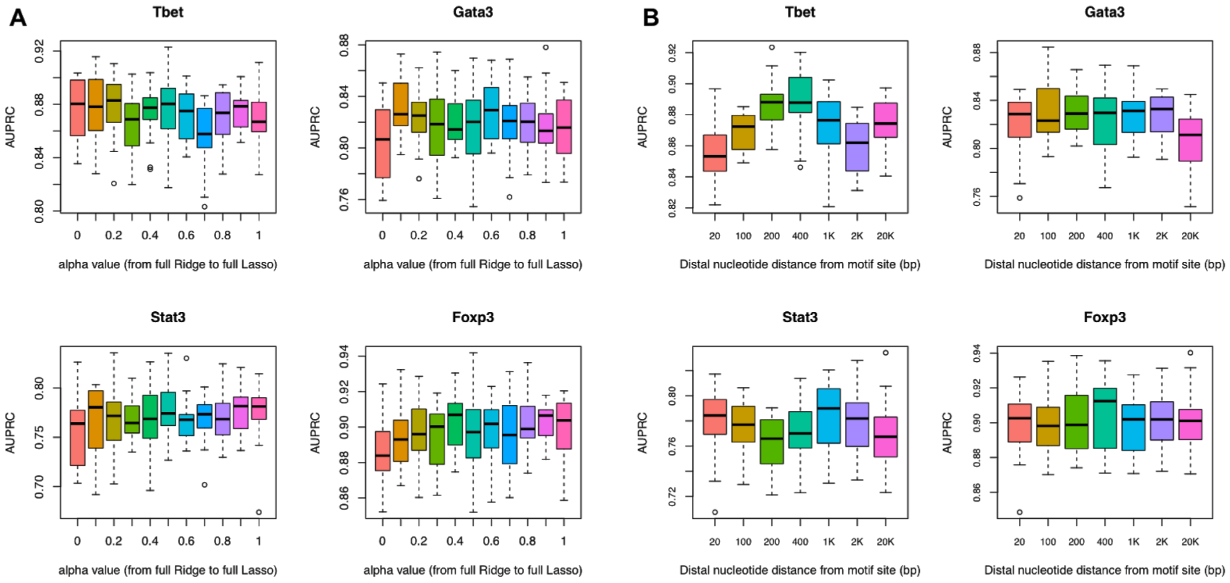


Figure A-1 legend. A) For a range of values of alpha, alpha represents the weight on the lasso penalty and 1-alpha represents the weight on the ridge penalty, we computed AUPRCs over 50 instances of running IMPACT for four canonical CD4+ T cell TFs. No value of alpha significantly outperformed the others. B) For a range of parameters characterizing how far away from the motif site we additionally check for feature overlap, we computed AUPRCs over 50 instances of running IMPACT for the same four CD4+ T cell TFs. No parameter value significantly outperformed the others.

Figure A-2. Effect of titration of TF motif detection threshold on training data selection and genomic annotation

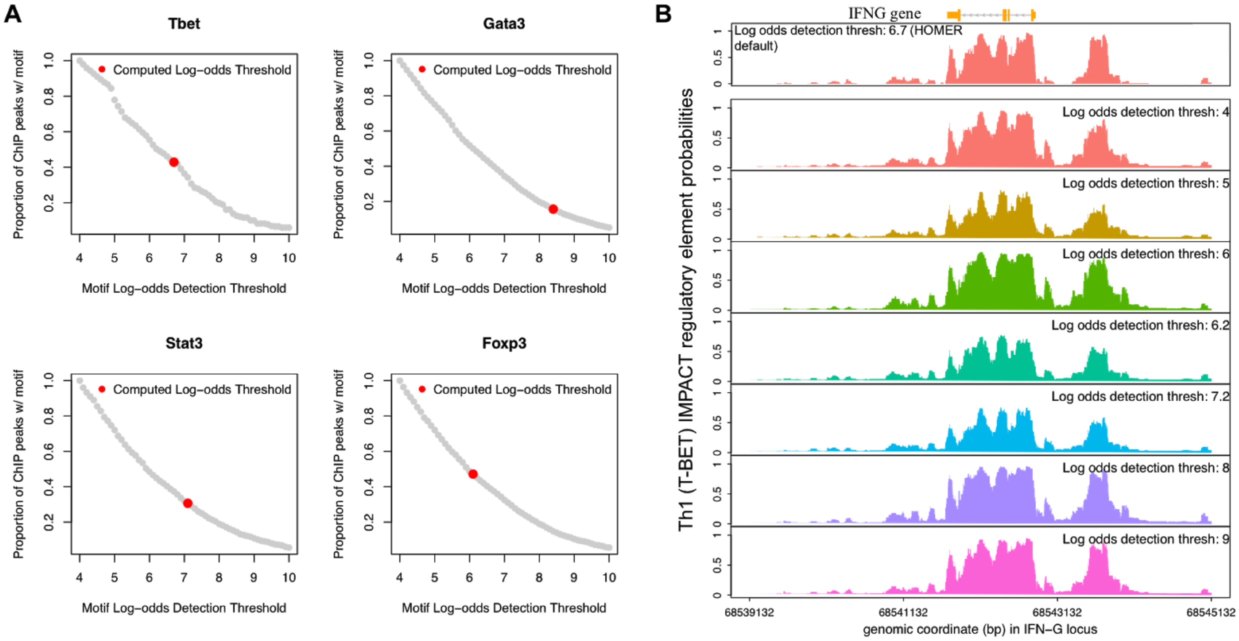


Figure A-2 legend. (A) Proportion of ChIP-seq peaks per CD4+ T cell TF with a detectable TF-specific motif over a range of log-odds detection thresholds from most lenient (left) to strictest (right). We indicate in red the optimized default HOMER log-odds detection threshold. (B) Genomic annotation with Th1 (T-BET) IMPACT around the IFN-G locus on chromosome 12 as a function of the motif detection threshold used to select the training data.

Figure A-3. Comparing IMPACT TF binding prediction performance using different feature characterization and gold standard characterization strategies

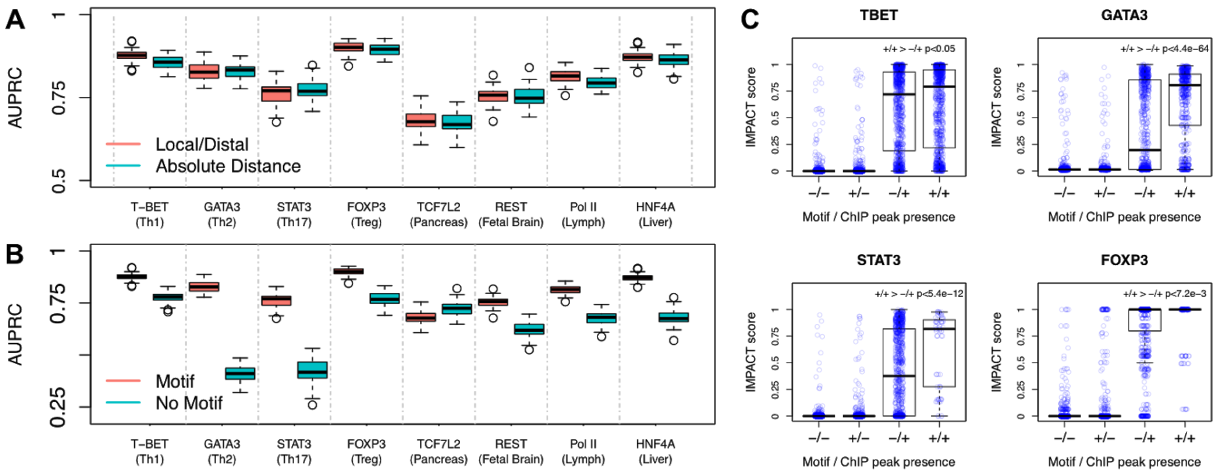


Figure A-3 legend. (A) We computed distributions of AUPRCs over 50 trials across 8 TFs to test the difference in TF binding prediction performance achieved by two feature representations. “Local/Distal” indicates the binary feature characterization, assessing feature overlap directly at the motif site and additionally at a more distal nucleotide, both upstream and downstream. “Absolute Distance” indicates a continuous and completely non-sparse feature characterization, in which the distance between each motif site and each feature is computed. (B) We computed distributions of AUPRCs over 50 trials across 8 TFs to test the difference in TF binding prediction performance achieved by different gold standard representations. “Motif” indicates that we pruned ChIP-seq peaks for motif sites and use these as gold standard bound or unbound regions. “No Motif” indicates that we used entire ChIP-seq peaks as the gold standard bound regions and permuted these regions genome-wide to obtain unbound regions. (C) IMPACT regulatory element probabilities are significantly higher at regions containing both a motif site and a ChIP-seq peak compared to ChIP-seq peak alone.

Figure A-4. IMPACT cell-state-specific regulatory element probabilities of the TF-bound motif sites and TF-unbound motif sites.

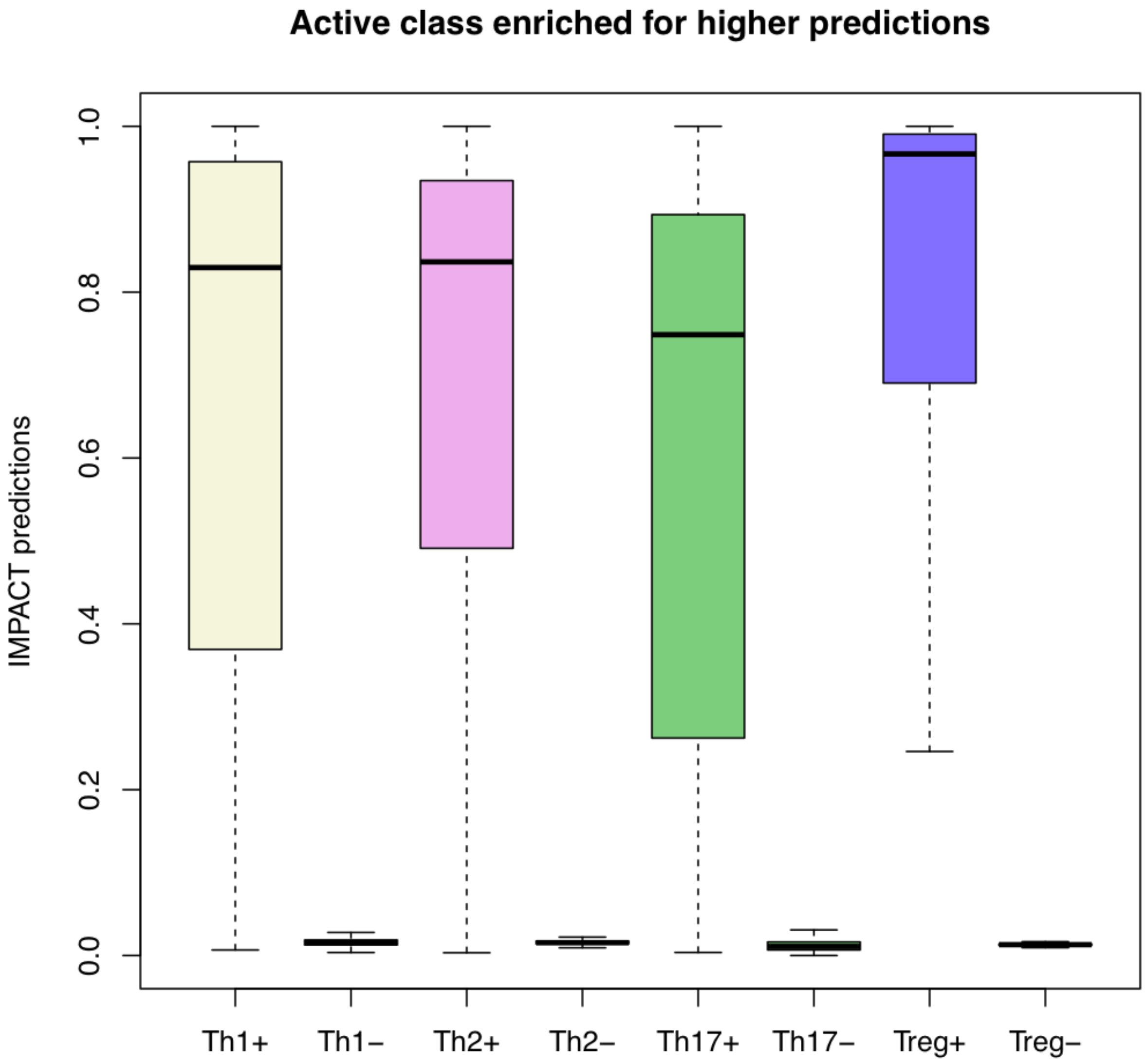


Figure A-4 legend. IMPACT regulatory predictions are significantly higher for actively bound motif sites, evidenced by ChIP-seq, than unbound motifs genome-wide. This demonstrates that IMPACT clearly distinguishes between regulatory and inactive/non-specific regulatory regions and assigns very low probabilities to the latter. Actively bound motif sites are represented by the modeled cell-state and a "+", whereas the unbound motif sites are represented by a "-".

Figure A-5. IMPACT TF binding performance assessment across 8 TFs.

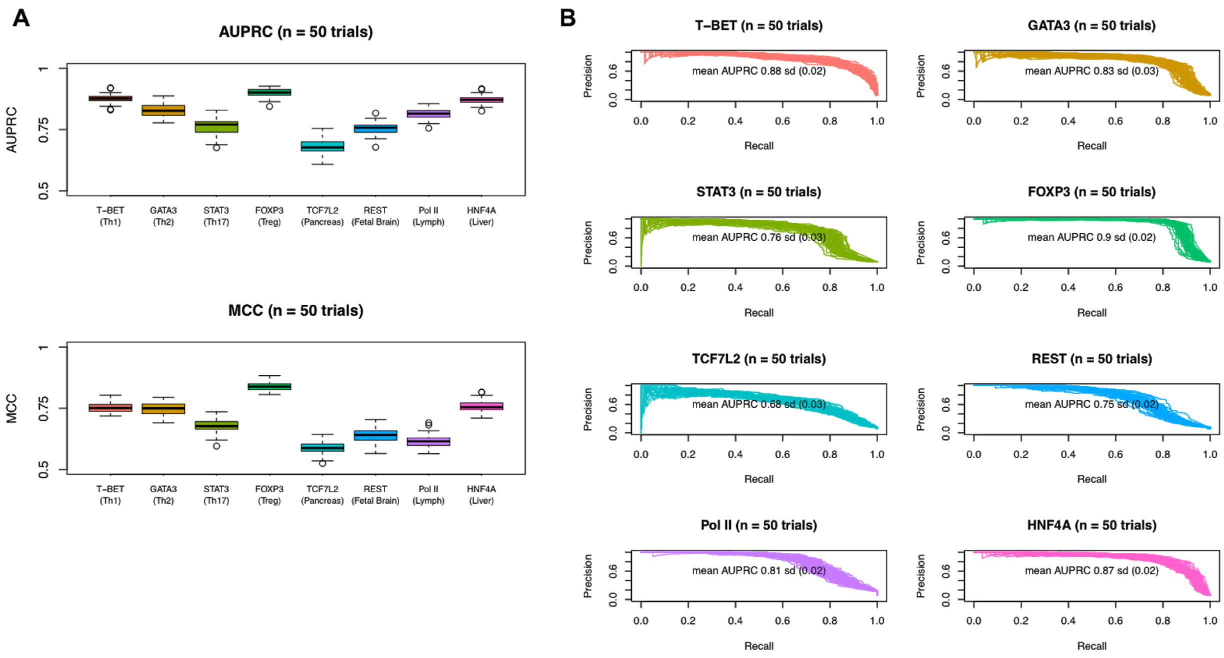


Figure A-5 legend. A) We use two metrics to illustrate IMPACT's TF binding predictive performance: area under the precision-recall curve (AUPRC) and Matthew's correlation coefficient (MCC); both metrics are appropriate for unbalanced class sets in binary classification problems. B) Here we plot the precision-recall curves for each of 8 TFs and note the mean AUPROC and standard deviation in parentheses.

Figure A-6. IMPACT benchmarking against TF binding prediction methods MocapG, MocapS, and Virtual ChIP-seq.

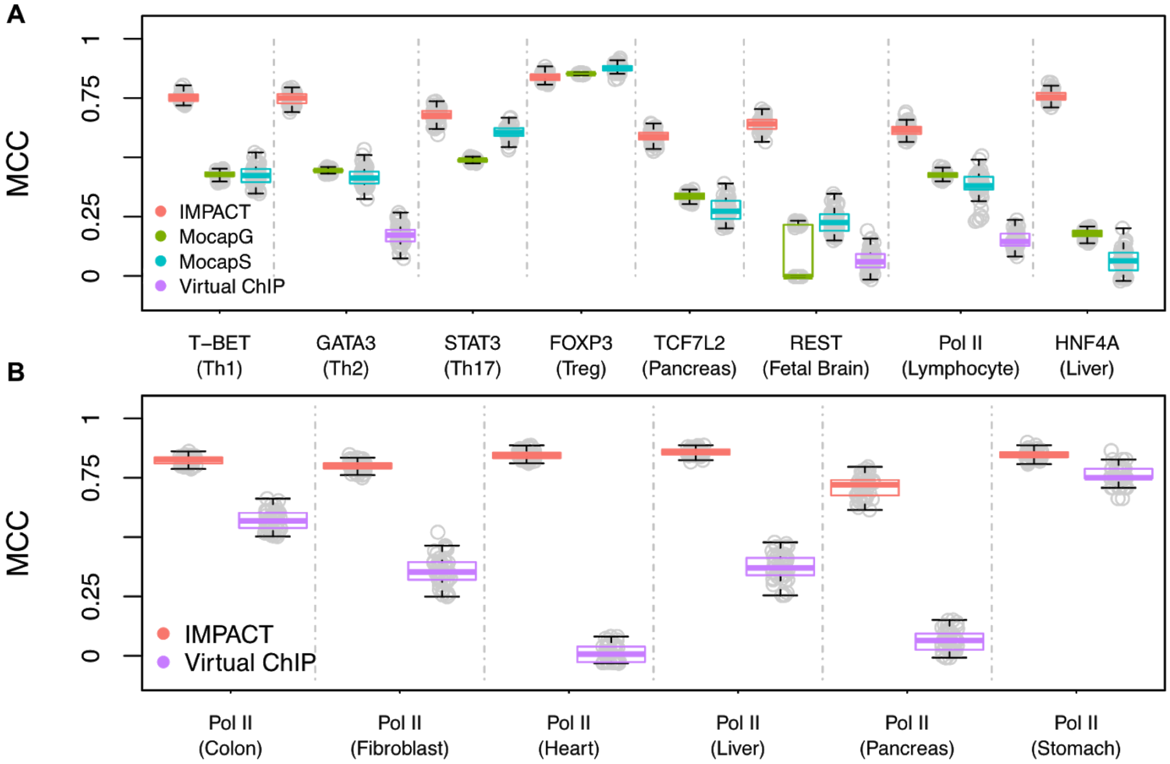


Figure A-6 legend: We additionally evaluated the TF binding predictive performance using Matthew's correlation coefficient. (A) We compared TF binding prediction across 8 TFs, requiring that different methods use consistent training and test data. (B) To further benchmark against Virtual ChIP-seq, we compared cell-type-specific TF binding prediction of RNA Pol II across 6 additional cell types.

Figure A-7. Cell-state-specific IMPACT predictions for canonical target genes of the four key CD4+ T cell TFs; IFNG (T-BET), IL4 (GATA3), SOCS3 (STAT3), and CTLA4 (FOXP3).

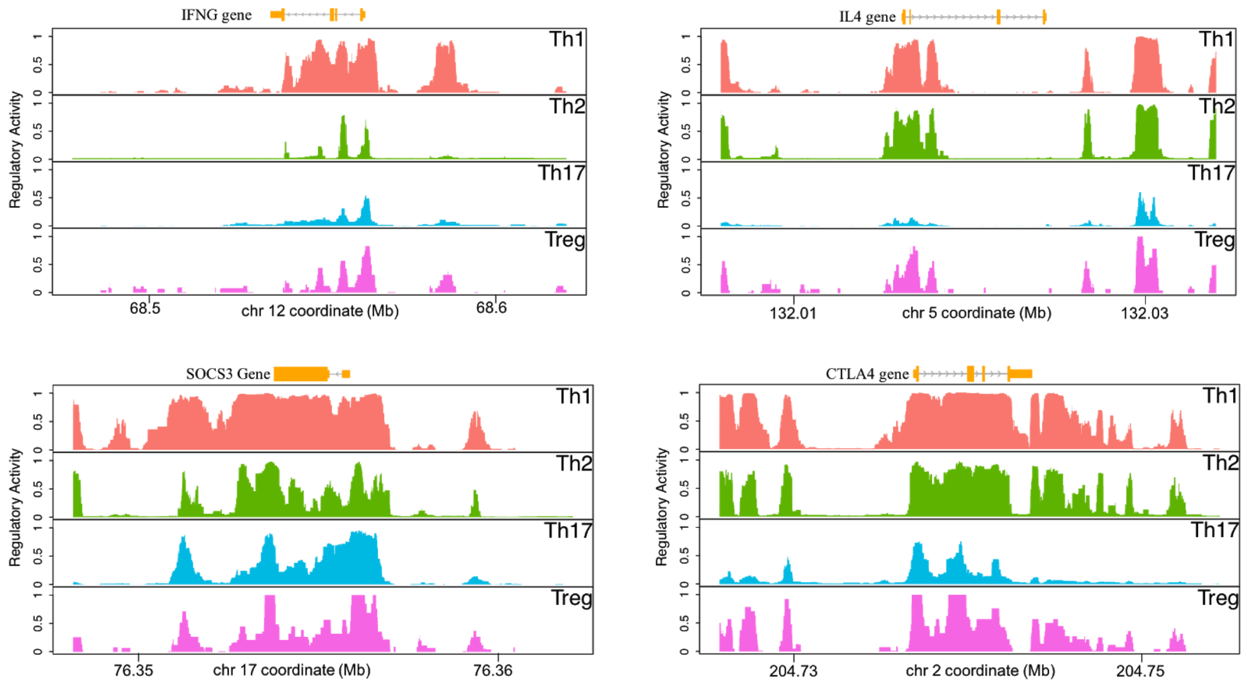


Figure A-7 Legend. We observe similar patterns of regulatory element probabilities across several cell-states for the same gene; this could be due the interrelatedness and co-regulation of CD4+ T cell immune programs. While we observe generally shared patterns across genes, we anticipate finer differences at the variant level, due to the differences in quantitative epigenomic profiles.

Figure A-8. Representative elastic net logistic regression coefficients () for IMPACT features in four CD4+ T cell-states with two different gold standard characterizations.

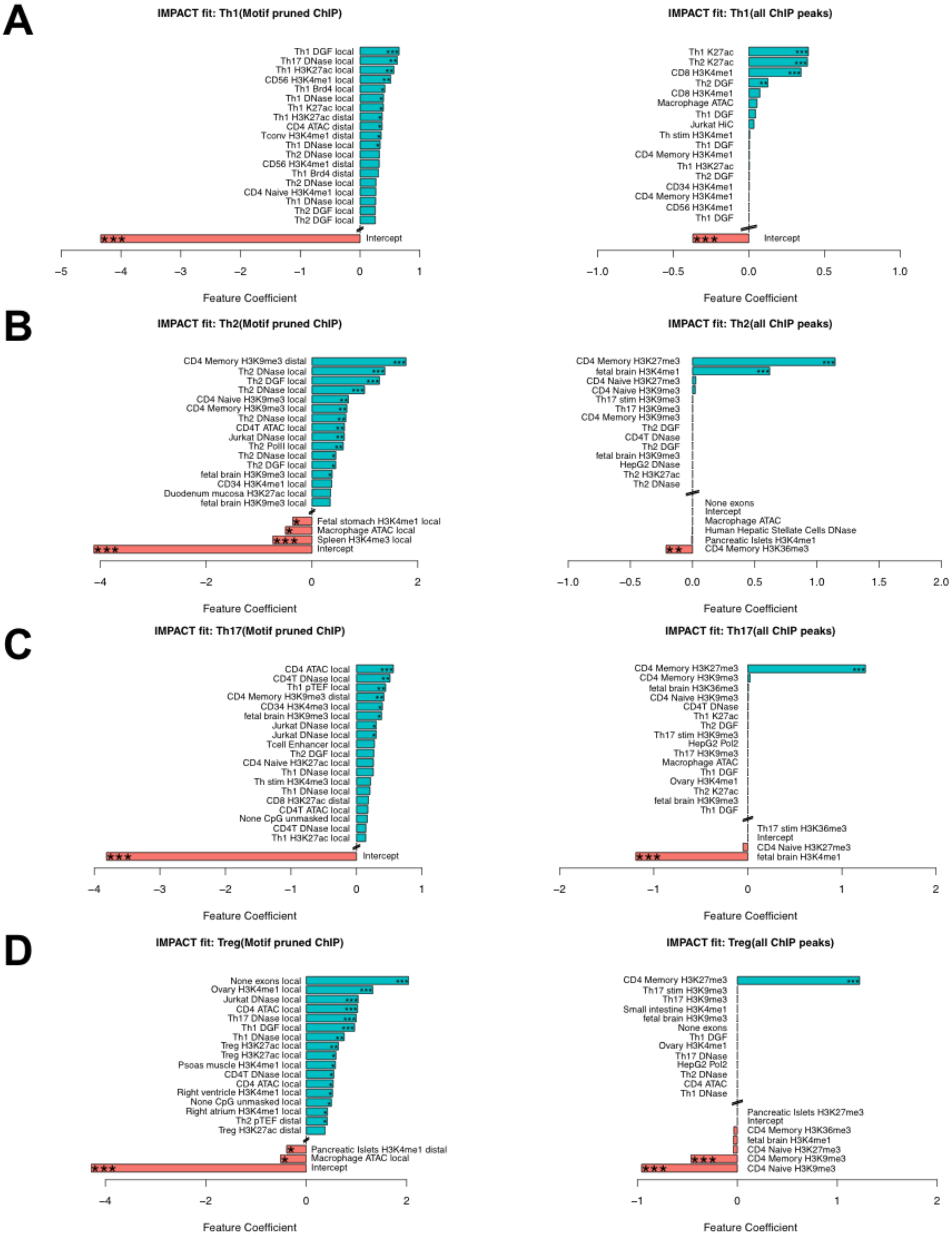

A
IMPACT fit: Th1(Motif pruned ChIP)
Th1 DGF local
Th17 DNase local
Th1 H3K27ac local
CD56 H3K4me1 local
Th1 Brd4 local
Th1 DNase local
Th1 K27ac local
Th1 H3K27ac distal
CD4 ATAC distal
Tconv H3K4me1 distal
Th1 DNase local
Th2 DNase local
CD56 H3K4me1 distal
Th1 Brd4 distal
Th2 DNase local
CD4 Naive H3K4me1 local
Th1 DNase local
Th2 DGF local
Th2 DGF local
Intercept
−5
−4
−3
−2
−1
0
1
Feature Coefficient
IMPACT fit: Th1(all ChIP peaks)
Th1 K27ac
Th2 K27ac
CD8 H3K4me1
Th2 DGF
CD8 H3K4me1
Macrophage ATAC
Th1 DGF
Jurkat HiC
Th stim H3K4me1
Th1 DGF
CD4 Memory H3K4me1
Th1 H3K27ac
Th2 DGF
CD34 H3K4me1
CD4 Memory H3K4me1
CD56 H3K4me1
Th1 DGF
Intercept
−1.0
−0.5
0.0
0.5
1.0
Feature Coefficient
B
IMPACT fit: Th2(Motif pruned ChIP)
CD4 Memory H3K9me3 distal
Th2 DNase local
Th2 DGF local
Th2 DNase local
CD4 Naive H3K9me3 local
CD4 Memory H3K9me3 local
Th2 DNase local
CD4T ATAC local
Jurkat DNase local
Th2 PolII local
Th2 DNase local
Th2 DGF local
fetal brain H3K9me3 local
CD34 H3K4me1 local
Duodenum mucosa H3K27ac local
fetal brain H3K9me3 local
Fetal stomach H3K4me1 local
Macrophage ATAC local
Spleen H3K4me3 local
Intercept
−4
−2
0
2
Feature Coefficient
IMPACT fit: Th2(all ChIP peaks)
CD4 Memory H3K27me3
fetal brain H3K4me1
CD4 Naive H3K27me3
CD4 Naive H3K9me3
Th17 stim H3K9me3
Th17 H3K9me3
CD4 Memory H3K9me3
Th2 DGF
CD4T DNase
Th2 DGF
fetal brain H3K9me3
HepG2 DNase
Th2 H3K27ac
Th2 DNase
None exons
Intercept
Macrophage ATAC
Human Hepatic Stellate Cells DNase
Pancreatic Islets H3K4me1
CD4 Memory H3K36me3
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Feature Coefficient
C
IMPACT fit: Th17(Motif pruned ChIP)
CD4 ATAC local
CD4T DNase local
Th1 pTEF local
CD4 Memory H3K9me3 distal
CD34 H3K4me3 local
fetal brain H3K9me3 local
Jurkat DNase local
Jurkat DNase local
Tcell Enhancer local
Th2 DGF local
CD4 Naive H3K27ac local
Th1 DNase local
Th stim H3K4me3 local
Th1 DNase local
CD8 H3K27ac distal
CD4T ATAC local
None CpG unmasked local
CD4T DNase local
Th1 H3K27ac local
Intercept
−4
−3
−2
−1
0
1
Feature Coefficient
IMPACT fit: Th17(all ChIP peaks)
CD4 Memory H3K27me3
CD4 Memory H3K9me3
fetal brain H3K36me3
CD4 Naive H3K9me3
CD4T DNase
Th1 K27ac
Th2 DGF
Th17 stim H3K9me3
HepG2 Pol2
Th17 H3K9me3
Macrophage ATAC
Th1 DGF
Ovary H3K4me1
Th2 K27ac
fetal brain H3K9me3
Th1 DGF
Th17 stim H3K36me3
Intercept
CD4 Naive H3K27me3
fetal brain H3K4me1
−2
−1
0
1
2
Feature Coefficient
D
IMPACT fit: Treg(Motif pruned ChIP)
None exons local
Ovary H3K4me1 local
Jurkat DNase local
CD4 ATAC local
Th17 DNase local
Th1 DGF local
Th1 DNase local
Treg H3K27ac local
Treg H3K27ac local
Psoas muscle H3K4me1 local
CD4T DNase local
CD4 ATAC local
Right ventricle H3K4me1 local
None CpG unmasked local
Right atrium H3K4me1 local
Th2 pTEF distal
Treg H3K27ac distal
Pancreatic Islets H3K4me1 distal
Macrophage ATAC local
Intercept
−4
−2
0
2
Feature Coefficient
IMPACT fit: Treg(all ChIP peaks)
CD4 Memory H3K27me3
Th17 stim H3K9me3
Th17 H3K9me3
Small intestine H3K4me1
fetal brain H3K9me3
None exons
Th1 DGF
Ovary H3K4me1
Th17 DNase
HepG2 Pol2
Th2 DNase
CD4 ATAC
Th1 DNase
Pancreatic Islets H3K27me3
Intercept
CD4 Memory H3K36me3
fetal brain H3K4me1
CD4 Naive H3K27me3
CD4 Memory H3K9me3
CD4 Naive H3K9me3
−1
0
1
2
Feature Coefficient

Figure A-8 Legend. Here we illustrate the most representative IMPACT features achieving a β deviating at least 1 sd in magnitude from the mean across 515 epigenomic and sequence-based features. Significantly positive β (green) means the feature is indicative of TF binding, whereas significantly negative β (red) means the feature is not indicative of TF binding. (A-D) IMPACT models trained on TF ChIP-seq from canonical CD4+ T cell-state TFs. (Left) IMPACT models trained on gold standard motif sites. (Right) IMPACT models trained on entire ChIP-seq peaks with no motif information. *** indicates |β| is at least 3 standard deviations (sd) from the mean β, ** indicates at least 2, and * indicates at least 1.

Figure A-9. 2D hierarchical clustering of pairwise Pearson r correlation values between CD4+ T cell-state IMPACT annotations and the 45 most strongly correlated features for each cell-state.

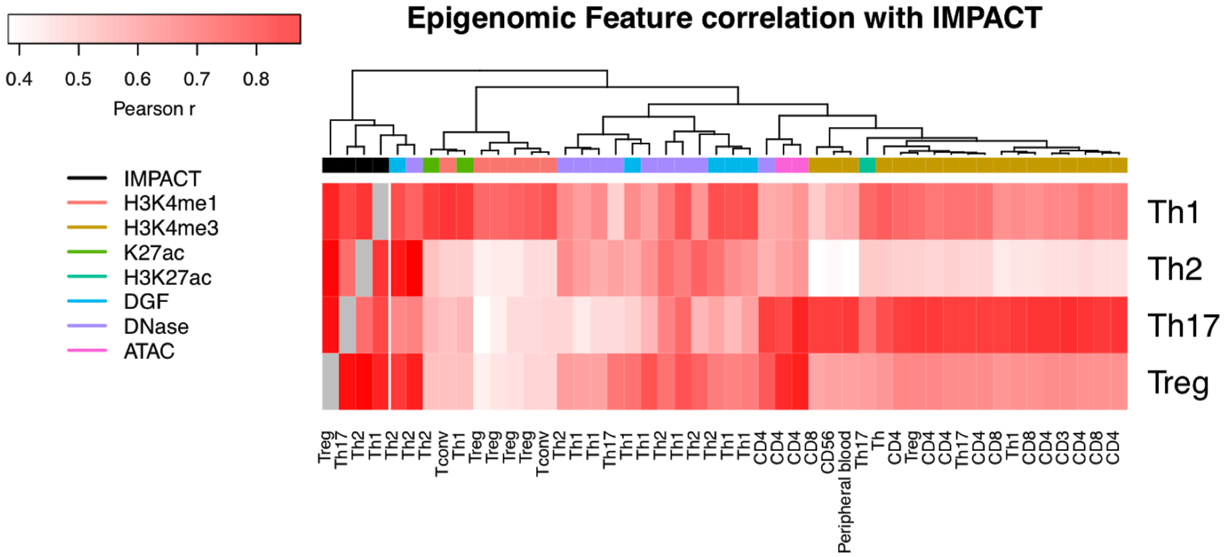


Figure A-9 legend. IMPACT annotations are highly correlated and particular cell-states are associated with categories of features, such as Th17 and H3K4me3 tracks and Th1 and H3K4me1 tracks. Each feature annotation is represented by a vector of length M, common variants (MAF ≥ 0.05), in which a value of 1 indicates intersection with a feature and a value of 0 indicates no intersection. In the case of

IMPACT, each cell-state is represented by a vector of length M, where each variant is assigned a continuous value representing the cell-state-specific regulatory element probability.

Figure A-10. Evaluating IMPACT feature category importance.

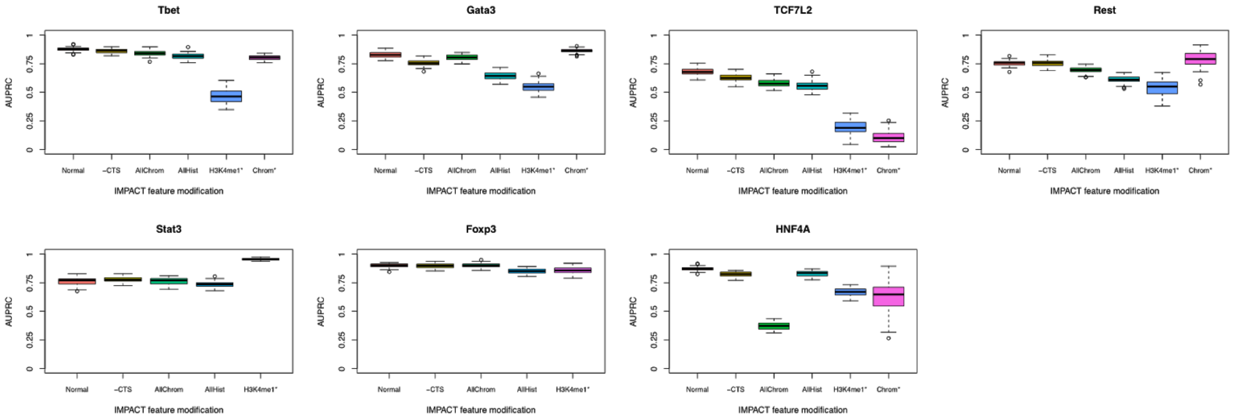


Figure A-10 legend. We evaluated CD4+ T cell-state IMPACT annotations while removing categories of features to assess the importance and redundancy of that category amongst other features. From left to right we show distributions of the AUROC over 50 trials for the unaltered IMPACT features (“Normal”), removal of cell-state-specific features (“-CTS”), inclusion of any chromatin annotation (“AllChrom”), inclusion of any histone modification annotation (“AllHist”), inclusion of only cell-state-specific H3K4me1 (“H3K4me1*”), and inclusion of only cell-state-specific open chromatin if applicable (“Chrom*”).

Figure A-11. eQTL chi-squared enrichments for Pol II IMPACT in other cell types.

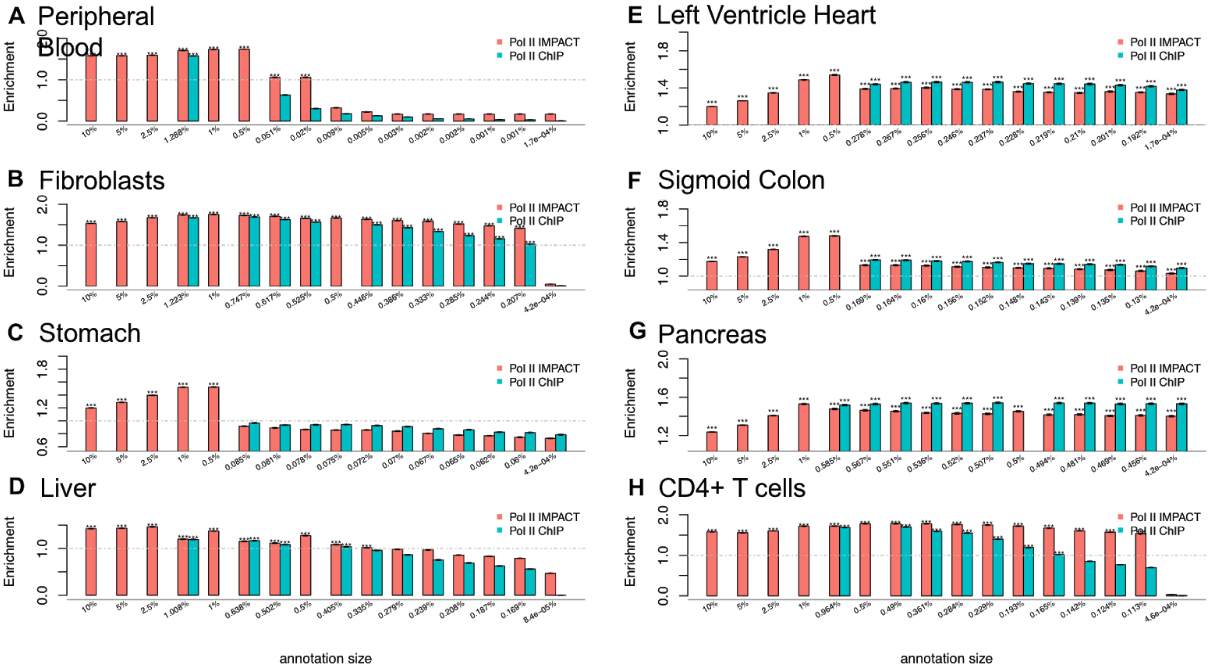


Figure A-11 legend. We computed the enrichment of cis eQTL chi-squared association values in Pol II IMPACT annotations and Pol II ChIP-seq annotations, created for peripheral blood (A), fibroblasts (B), stomach (C), liver (D), left ventricle heart (E), sigmoid colon (F), pancreas (G), and CD4+ T cells (H). We compared the enrichments of Pol II IMPACT across a range of annotation cutoffs and against Pol II ChIP-seq used to train the IMPACT model, also over a range of peak significance cutoffs. Annotation size is listed on the x-axis and *** indicates permutation *P* < 0.001. Intervals at the top of each bar represent the 95% confidence interval of the enrichment estimate.

Figure A-12. S-LDSC effect size analysis of IMPACT annotation with respect to RA.

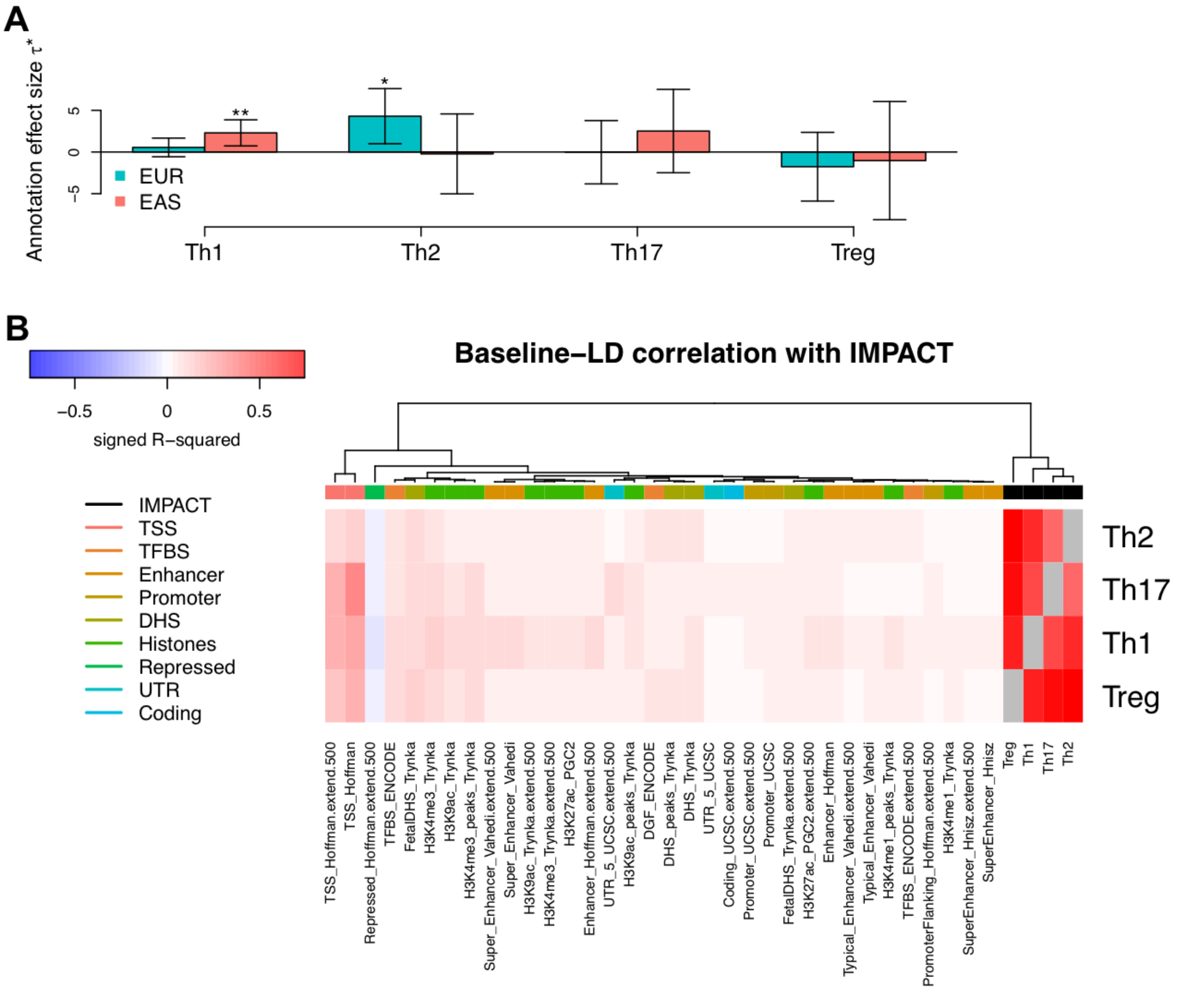


Figure A-12 legend. A) Per-annotation standardized effect sizes (τ *) for CD4+ T IMPACT annotations in RA, while conditioned on each other and the baseline-LD annotations (1 S-LDSC model per population). In the European model (green), τ * of Th2 IMPACT is significantly positive. In the East Asian model (red), τ * of Th1 IMPACT is significantly positive. We interpret this to mean that Th2 and Th1 IMPACT regulatory element probabilities are most correlated with per-SNP RA h2 and better capture the bulk of the RA polygenic signal. B) 2D hierarchical clustering of pairwise signed Pearson R-squared correlations between CD4+ IMPACT annotations and 34 most strongly correlated baseline-LD annotations.

Figure A-13. S-LDSC RA heritability enrichments of IMPACT annotations and other CD4+ T cell annotations from experimental assays.

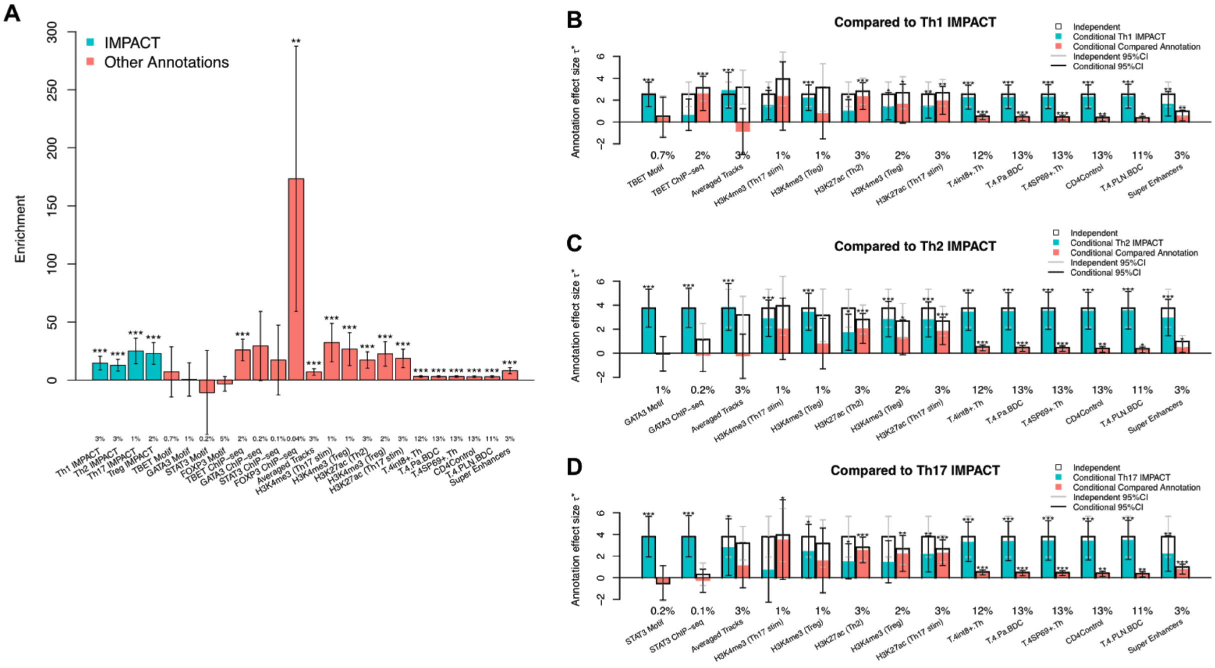


Figure A-13 legend. a) Enrichment of CD4+ IMPACT annotations compared to other T cell related functional annotations. b) τ * values for pre and post conditioning of IMPACT annotations (b-d: Th1, Th2, and Th17) on other T cell related functional annotations. Only the Th2 annotation has consistently positive τ * values while conditioned on other annotations, although H3K27ac in Th2 capture more RA h2. T-BET ChIP-seq and H3K27ac in Th2 outperforms the Th1 annotation and several annotations outperform the Th17 annotation in capturing RA h2. For panels a and b, no asterisk denotes *P* < 0.05, 1 asterisk *P* < 0.05, 2 asterisks *P* < 0.01, 3 asterisks *P* < 0.001.

Figure A-14. IMPACT cell-state-specific regulatory element predictions at RA-associated loci on chromosomes 1, 2, and 3.

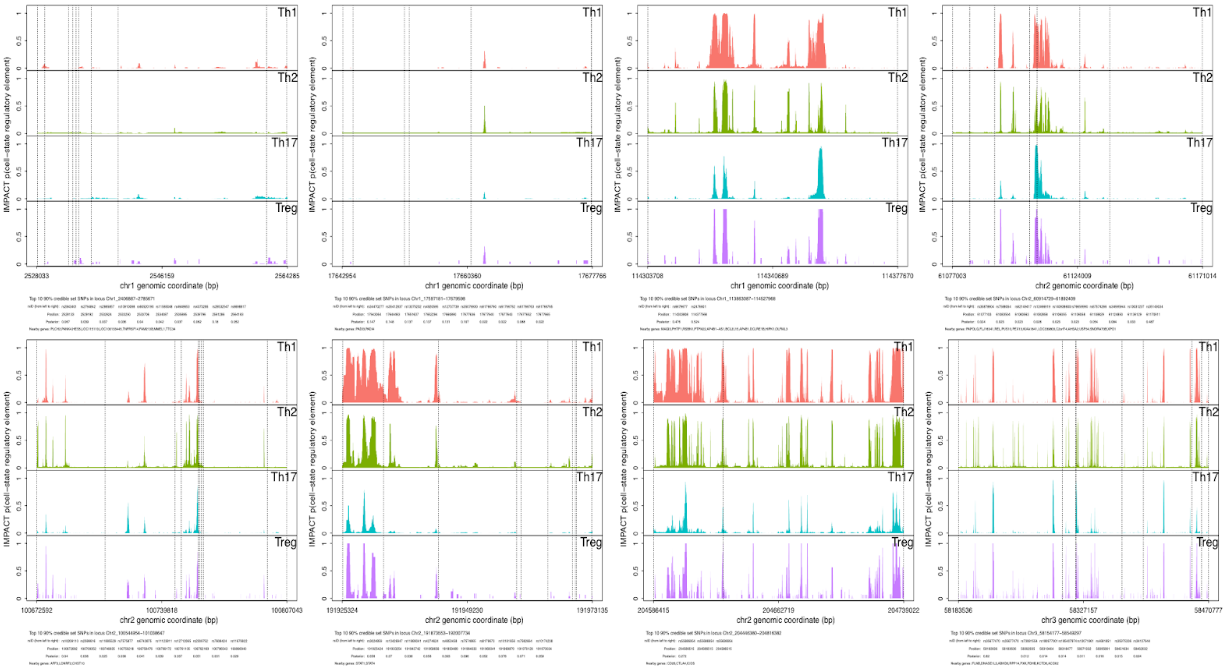


Figure A-14 legend. Intersection of nucleotide-resolution CD4+ T cell-state IMPACT annotations with putatively causal variants. We note the position of the top 10, if applicable, 90% credible set SNPs, their rsIDs, and the nearest genes.

Figure A-15. IMPACT cell-state-specific regulatory element predictions at RA-associated loci on chromosomes 4, 5, 6, 7, and 15.

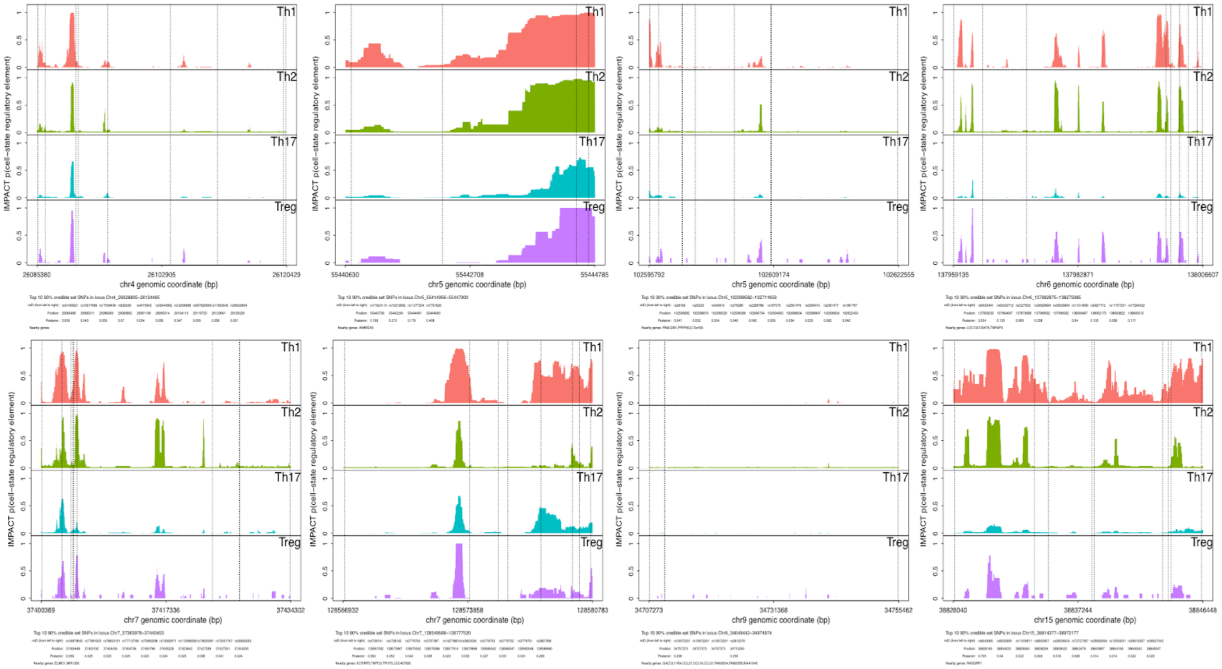


Figure A-15 legend. Intersection of nucleotide-resolution CD4+ T cell-state IMPACT annotations with putatively causal variants. We note the position of the top 10, if applicable, 90% credible set SNPs, their rsIDs, and the nearest genes.

Figure A-16. IMPACT cell-state-specific regulatory element predictions at RA-associated loci on chromosomes 17, 18 and 19.

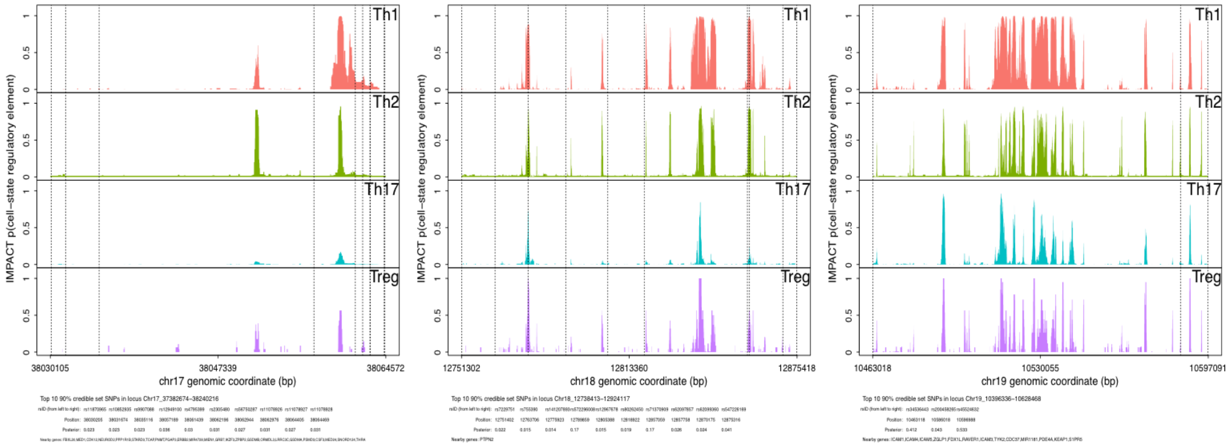


Figure A-16 legend. Intersection of nucleotide-resolution CD4+ T cell-state IMPACT annotations with putatively causal variants. We note the position of the top 10, if applicable, 90% credible set SNPs, their rsIDs, and the nearest genes.

Figure A-17. Effect of annotation size on S-LDSC heritability enrichment estimates for the 5 most enriched CD4+ T cell specifically expressed gene sets.

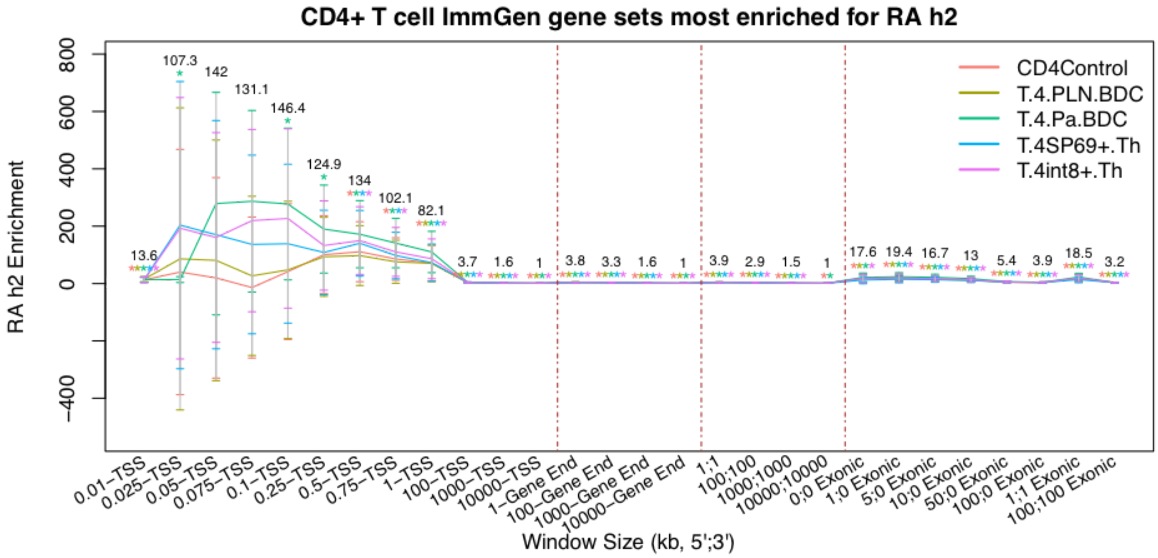


Figure A-17 legend. For multiple window sizes around each gene, we computed the heritability enrichment estimate for RA for 5 different CD4+ T cell specifically expressed gene sets, where each gene set contains approximately 1,000 genes.

Appendix B

Supplementary Information for Chapter 3

Significant IMPACT annotation-trait associations
We identified at least one statistically significant IMPACT annotation association with 95 of 111 polygenic traits. These 95 account for 60 of 69 European phenotypes and 35 of 42 East Asian phenotypes. Analogously, across 707 cell type regulatory annotations, we identified at least one significant annotation-trait association for 566 annotations at 5% FDR. For all trait-annotation pairs, the computed τ * and enrichment estimates, along with their standard errors can be found in **Tables B-4-8**.

Annotations and traits with no observed heritability enrichment
For 16 polygenic traits, we observed no statistically significant annotation association. Of these 16 polygenic traits, 9 were from European GWAS; these are anorexia, cataract, “ever smoked”, three pigmentation phenotypes (skin, sunburn, tanning), and three heart disease phenotypes (CHF, IS, AF). The remaining 7 traits with no annotation associations from East Asian GWAS were cataract, COPD, IS, keloid, osteoporosis, pancreatic cancer, and pollinosis. Likewise, for 141 IMPACT annotations, we observed no statistically significant trait association. These annotations included melanoma and heart-labeled annotations (**Figure B-29**). Just over 40% of sarcoma annotations were significantly associated with at least one trait; for all other tissue types, more than 60% of the corresponding annotations were significantly associated with at least one trait. We found that number of training ChIP-seq peaks were significantly correlated with both the size of annotation and the AUPRC of the TF binding model (Pearson $r = 0.22$, $P < 1.5\text{e-}9$; Pearson $r = 0.39$, $P < 1.5\text{e-}24$, respectively) (**Figure B-29**). However, the AUPRC and size of annotation are significantly negatively correlated (Pearson $r = -0.25$, $P < 4.8\text{e-}11$). This perhaps indicates that models with a small number of training peaks and above-average AUPRC (overfitting) will lead to smaller annotations which don’t adequately cover the polygenic space, leading to fewer significant heritability enrichments. Moreover, we found that these unassociated annotations have generally significantly smaller annotation sizes ($P < 7.0\text{e-}10$), significantly higher TF binding model AUPRCs ($P < 3.2\text{e-}18$), significantly less training data ($P < 0.03$), and are biased for particular cell types (**Figure B-29**).

Deep learning comparison across 69 EUR traits
As we performed a more thorough comparison of heritability captured by IMPACT compared to deep learning annotations among the five representative traits by collecting 123 relevant

annotations, such an analysis was challenging to perform across all 69 EUR traits. As Basenji and DeepSEA annotations from a previous study[109] accounted for the lead annotation among the five representative traits, we applied these 32 annotations to partition the heritability of the remaining 64 EUR traits. We found that IMPACT annotations captured more heritability (49.5%, sem = 3.3%) than both lead Basenji deep learning annotations (31.9%, sem = 1.9%, one-tailed paired wilcoxon $P < 2.0e-11$) (**Figure B-8, Table B-13**) and lead DeepSEA deep learning annotations (27.5%, sem = 1.2%, one-tailed paired wilcoxon $P < 1.4e-10$) (**Figure B-8, Table B-13**). Moreover, the τ^* of lead IMPACT annotations was almost always greater than that reported for Basenji annotations (by a factor of 2.24x, one-tailed paired wilcoxon $P < 3.4e-11$) and for DeepSEA annotations (by a factor of 3.55x, one-tailed paired wilcoxon $P < 8.8e-12$, **Figure B-8, Table B-13**).

Regulatory concordance of complex traits

Not only did we observe shared regulatory biology between populations, but also among traits. Despite weak genetic correlation among different traits, we observed strong correlations of IMPACT annotation τ^* among traits, revealing large regulatory modules of immunity, white blood cell regulation, red blood cell (RBC) regulation, and body height (**Figure B-30**). These results suggest that while causal effects and variants may differ among biologically related traits, the regulatory elements in which these variants reside may be shared. Moreover, while genetic correlation approaches consider all genetic signals genome-wide which comprise true biological signal and artefact, we believe that IMPACT is more likely to identify true biological effects, which are shared between related traits, unlike artifactual signals.

Conditional S-LDSC analysis to identify independent annotation-trait associations

Before performing serial conditional analyses, for 9 polygenic traits, we observed a single associated cell type: EUR autism (breast), EAS breast cancer (breast), EAS cervical cancer (stem cell), EAS congestive heart failure (colon), EAS diastolic blood pressure (mesendoderm), EAS gastric cancer (stomach), EAS glaucoma (adipocytes), EAS systolic blood pressure (mesendoderm), EAS uterine fibroids (hematopoietic progenitors). However, for 86 traits, we observed that regulatory elements of multiple IMPACT annotations, mostly implicating diverse cell types, significantly capture heritability (**Figure B-9**). After performing serial conditional analyses to resolve dependent and independent associations, there remained a total of 142 independent cell type-trait associations (**Figure B-9**): 1 trait with 4 associations, 7 traits with 3, 30 traits with 2, 57 traits with 1, and 16 traits with none. Four annotations independently explained significant proportions of heritability in EUR prostate cancer: prostate (TFAP4), prostate (RUNX2), mesendoderm (PDX1), and cervix (NFYB). For seven European traits, three IMPACT annotations independently captured polygenic heritability: height (adipocytes, fibroblasts, lung), neutrophil count (monocytes, adipocytes, B cells), osteoporosis (myoblasts, mesenchymal stem cells, cervix), IBD (T cells and two B cell annotations), platelet count (PBMCs, hematopoietic progenitors, muscle), systolic blood pressure (endothelial, mesenchymal stem cells, fibroblasts), and white blood cell count (B cells, adipocytes, hematopoietic progenitors). For each of 22 European traits and 8 East Asian traits, we observed exactly two independent IMPACT annotation associations. Finally, for each of 30 European traits and 27 East Asian traits, we observed exactly one independent IMPACT association. For Crohn's (EUR), Th1s and naive CD4+ T cells independently captured heritability, suggesting two different biological mechanisms one via naive T cells and the other via memory effector cells. Although previous studies suggested an important role of T cells in UC[24], our study identified not only T cells but also B cells as contributors to disease pathogenesis. For UC (EUR), T cells and B cells contribute independently to explain heritability. In summary, we have elucidated the biology of some polygenic traits through resolving not only the most significantly associated cell type, but also secondary, tertiary, and quaternary independent mechanisms. These results also shed light on shared regulatory programs between cell types: in cases where prior to conditioning, we observed many diverse cell type associations, yet upon conditioning revealed a single independent signal. For example, in EUR RA, B cells were most strongly associated, while CD4+ memory T cell annotations also captured significant proportions of heritability. However, these T cell annotations were not associated independently of B cells, suggesting that RA heritability resides in shared regulatory elements between T and B cells. In summary, we have elucidated the biology of some polygenic traits through resolving not only the most significantly associated cell type, but also secondary, tertiary, and quaternary independent mechanisms.

To investigate the concordance of independent IMPACT signals across related traits, we considered clusters of functionally correlated traits from **Figure B-30**. Among the autoimmune disease and hematological trait cluster, encompassing eosinophil count, asthma, RA, and lymphocyte count, the CD4 T cell:BCL6 and Th1:TBX21 annotations were each three times listed as independent contributors. For the greater hematological trait cluster consisting of monocyte, neutrophil, white blood cell, basophil, platelet, lymphocyte, red blood cell counts as well as MCV, MCH, and MCHC, the PBMC:GATA1 annotation was eight times listed as an independent contributor. Lastly, for the endocrine cluster consisting of BMI, T2D, SBP, Hb, and Ht, the mesendoderm:PDX1 annotation was six times listed as an independent contributor. These observations reveal that there is indeed some degree of persistence of independent genetic contributors and may add a biological basis for the observed genetic correlations among these traits.

We note that our cell type interpretations above rely on the fidelity of the IMPACT model to accurately predict TF binding in the desired cell type; a poor model may learn an epigenetic signature that does not represent the desired cell type. The mean TF binding model AUPRC of independently associated IMPACT annotations was significantly less (mean AUPRC = 0.41, sem = 0.04) than than of all IMPACT annotations (mean AUPRC = 0.54, sem = 0.01, difference of means

P < 8.1e-4). This is consistent with our observation that IMPACT annotations with very high AUPRCs are less likely to capture polygenic heritability (**Figure B-29**).

Cell type composite annotations targeting multiple independent mechanisms of polygenic traits

In light of observing 38 phenotypes for which multiple cell type regulatory element annotations independently captured significant proportions of heritability, we created composite cell type annotations in hopes of improving heritability enrichments. For example, we observed that genetic variation governing neutrophil count (EUR) is independently accounted for by monocytes, adipocytes, and B cell regulatory elements. Then, we annotated SNPs genome-wide using a probabilistic OR gate as follows:

$$score_j = 1 - \prod_i^a (1 - IMPACT_{i,j}),$$

where *j* is the SNP index, *i* is the i^{th} annotation, *a* is the number of independently associated annotations for the trait of interest and $IMPACT_{i,j}$ is the IMPACT score of variant *j* in annotation *i*.

We created 38 composite cell type annotations and observed that these annotations captured significantly more overall enrichment (one-tailed paired wilcoxon *P* < 4.9e-10), significantly more per-SNP heritability in terms of τ * (one-tailed paired wilcoxon *P* < 3.2e-8), and significantly more heritability in the top 5% of SNPs (one-tailed paired wilcoxon *P* < 0.004) (**Figure B-31**).

Trends of multi-ethnic marginal effect size correlation at various *P* value thresholds

We observed that at lenient *P* value thresholds, the difference in correlation between EUR and EAS effect sizes is more pronounced using IMPACT annotations, suggesting that they may be more effective for prioritizing causal variation particularly when statistical evidence is weak. For example, at the most lenient *P* value thresholds between *P* < 1 and *P* > 3e-4, we observed more dramatic improvements in correlation using IMPACT while on the other hand, at more stringent *P* value thresholds, IMPACT annotations offer less of an improvement in multi-ethnic effect size correlation (**Figure B-32**).

Robustness of PRS analysis to scale on which effect sizes are estimated

For case/control diseases, we estimated marginal effect sizes on the logistic scale. To ensure that our results were consistent if effect sizes were to be estimated on the liability scale, for each of 5 case/control diseases considered in PRS analyses, we converted effect sizes from logistic scale to liability scale (**Material and Methods**). The conversion had negligible effects on our findings: 1) effect size estimates were nearly perfectly correlated (**Figure B-25**), 2) PRS values were also nearly perfectly correlated (**Figure B-26**), and 3) the predictive power of PRS models were highly consistent (for EUR PRS resulting in an average change in magnitude of pseudo-R^2 equivalent to 1.8e-5 or a 0.16% average increase in pseudo-R^2 values relative to logistic-based PRS; and for EAS PRS resulting in an average change in magnitude of pseudo-R^2 equivalent to 1.3e-4 or a 0.81% average increase in pseudo-R^2 values relative to logistic-based PRS, **Figure B-26**). These results demonstrate that the way in which effect sizes are defined has negligible effects on our findings.

Supplementary Tables

Supplementary Tables can be found in the online supplement of Amariuta* and Ishigaki* et al, *bioRxiv* 2020 and due to their size are not included below.

Supplementary Figures

Figure B-1. Comparison of IMPACT implementation from AJHG 2019 manuscript and current study.

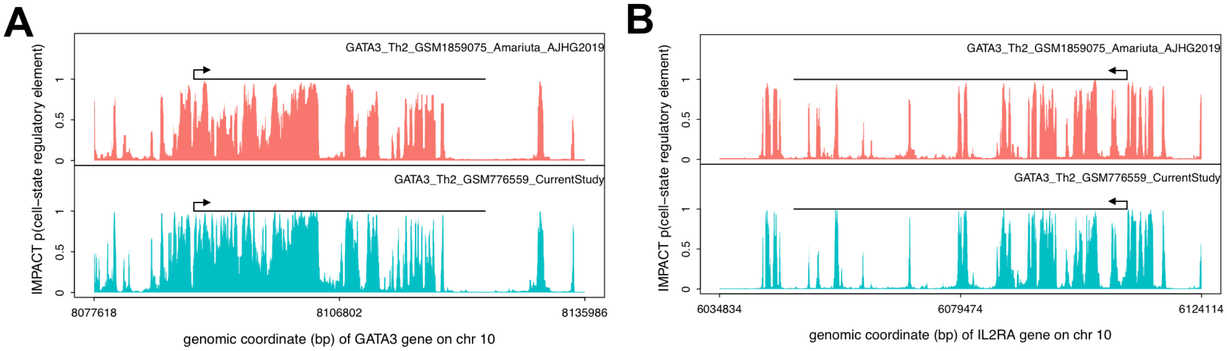


Figure B-1 legend. Consistency of IMPACT predictions for the same TF/cell type pair (GATA2/Th2) using different experiments and different feature sets: GSM1859075 used in Amariuta et al AJHG 2019 with 515 epigenetic features and GSM776559 used in the current study with 5,345 total epigenetic features. A) GATA3 gene locus on chr10. B) IL2RA gene locus on chr10.

Figure B-2. Overview of publicly available data used in study.

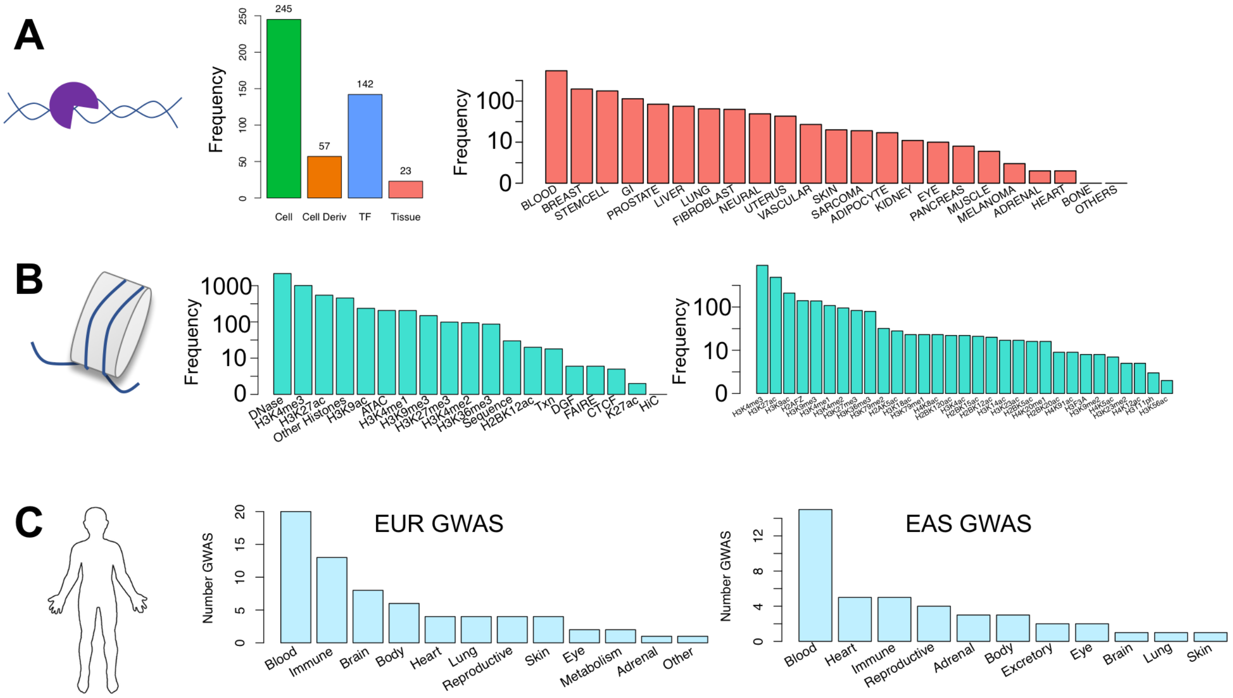
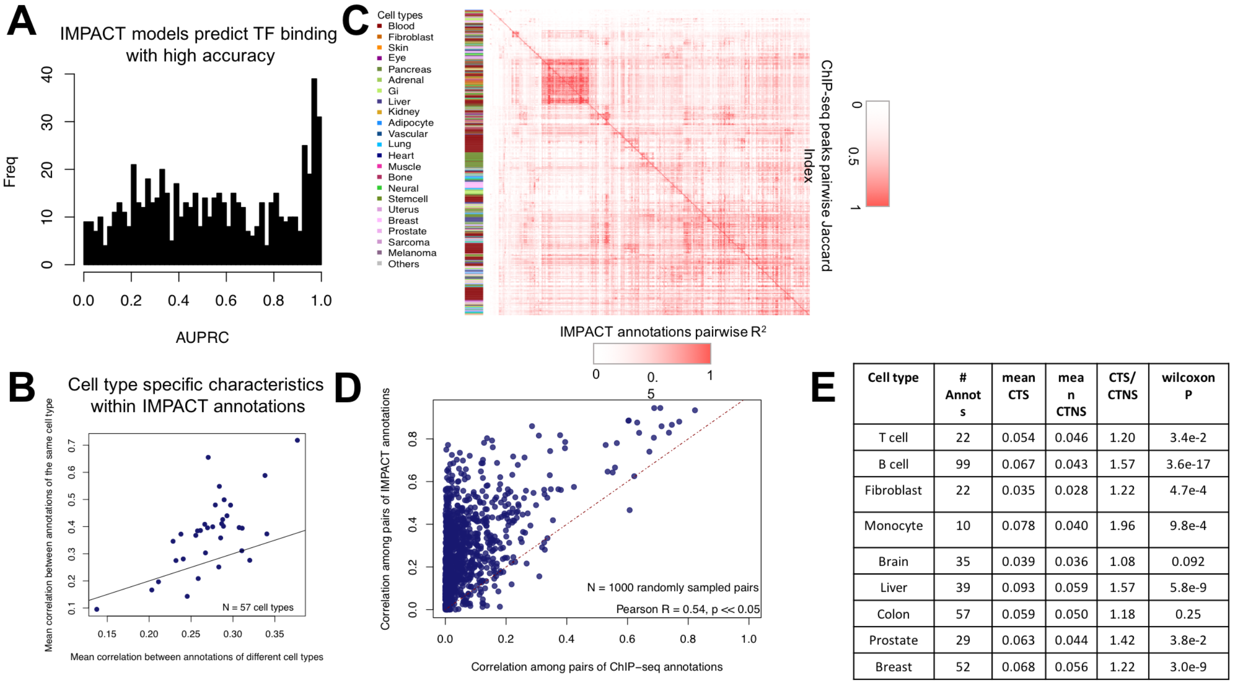


Figure B-2 legend. A) TF ChIP-seq collection from NCBI: (left) cell type and TF diversity where “Cell Deriv” indicates number of unique parental cell types, for example GM12878 and GM10847 are both B cell lines, (right) diversity of tissue types. B) (left) Epigenomic and sequence features to be used in IMPACT models, (right) diversity of histone modification ChIP-seq in features. C) Diversity of European (EUR) and East Asian (EAS) GWAS summary statistics across phenotypic categories.

Figure B-3. Summary of IMPACT-707 quality checks.



Cell type	# Annots	mean CTS	mean CTNS	CTS/ CTNS	wilcoxon P
T cell	22	0.054	0.046	1.20	3.4e-2
B cell	99	0.067	0.043	1.57	3.6e-17
Fibroblast	22	0.035	0.028	1.22	4.7e-4
Monocyte	10	0.078	0.040	1.96	9.8e-4
Brain	35	0.039	0.036	1.08	0.092
Liver	39	0.093	0.059	1.57	5.8e-9
Colon	57	0.059	0.050	1.18	0.25
Prostate	29	0.063	0.044	1.42	3.8e-2
Breast	52	0.068	0.056	1.22	3.0e-9

Figure B-3 legend. A) Histogram of prediction performance of 707 IMPACT models (metric = AUPRC). B) IMPACT annotations of the same cell type are more similar to one another than annotations of different cell types. C) Pairwise correlation of IMPACT regulatory element annotations (lower triangle of matrix) relative to pairwise correlation of corresponding TF ChIP-seq annotations (upper triangle of matrix). Pearson r was calculated using probabilities assigned to 779,355 SNPs on chr1 from phase 3 of 1000G (EUR), Jaccard indices were calculated for binary ChIP-seq tracks genome-wide, in which the size of the intersection of base pairs between two datasets was divided by the size of the union of base pairs. D) Pairwise correlations between 1000 randomly selected datasets between TF ChIP-seq and their corresponding IMPACT annotations; values sampled from C). E) IMPACT assigns larger cell-type-specific regulatory elements probabilities at cell-type-specifically expressed genes across nine cell types.

Figure B-4. Multi-ethnic genetic correlation and concordance of 707 IMPACT annotations with the baseline LD annotations.

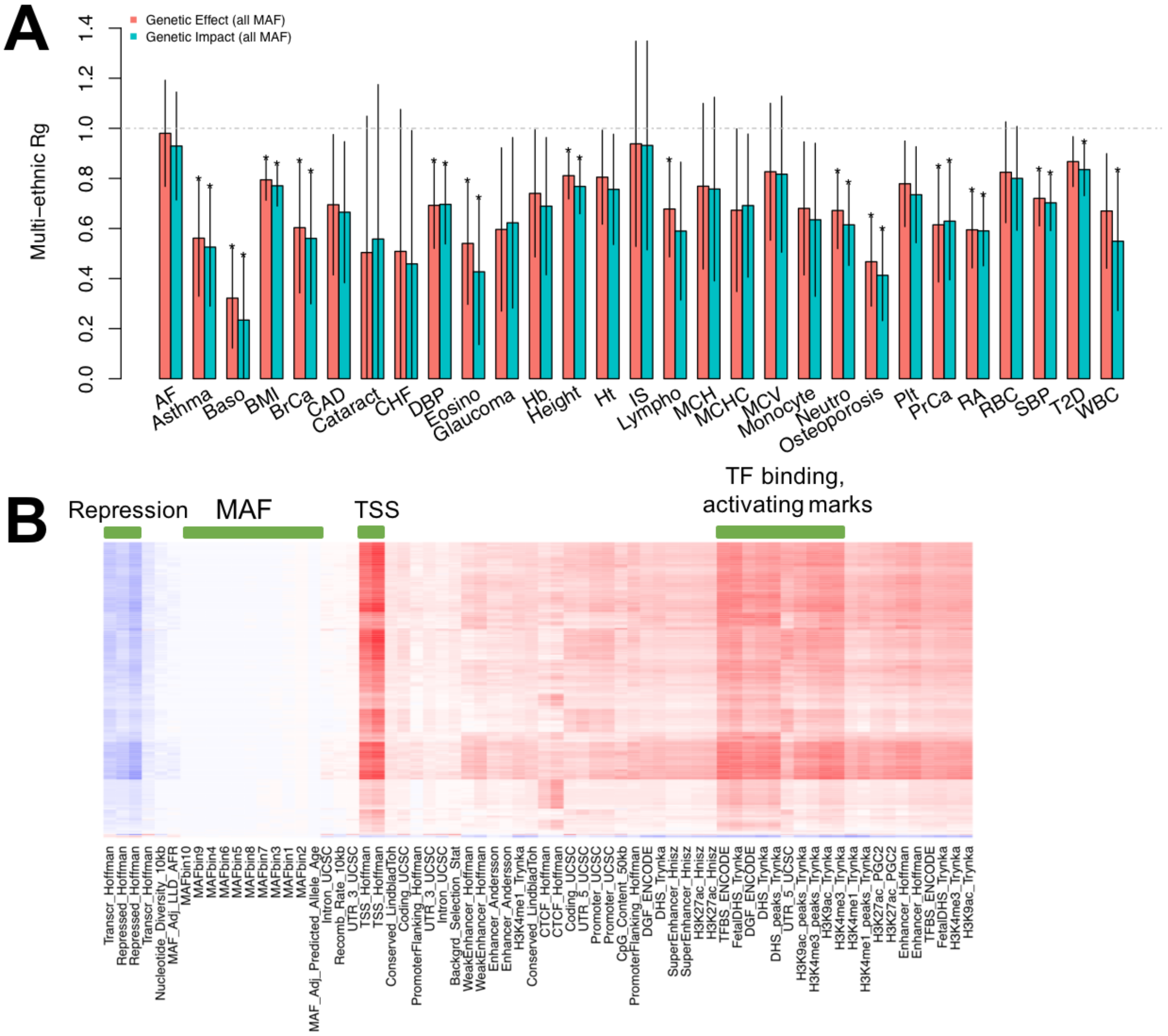


Figure B-4 legend. A) For the 29 traits for which we collected both EUR and EAS GWAS summary statistics, we computed the multi-ethnic genetic correlation with Popcorn. According to genetic effect, for 13 traits, the genetic correlation is significantly less than 1, indicated by an asterisk (*P* < 0.05 / 29 traits). We plot both the genetic correlation computed separately using genetic effect (effect size estimates unnormalized to allele frequency) and genetic impact (allele variance normalized effect sizes). B) IMPACT annotations correlate most with TSS, TFBS, and activation histone mark annotations, while no correlation is present with European ancestry MAF bins.

Figure B-5. Heritability captured by top 5% of SNPs according to lead IMPACT annotations per trait.

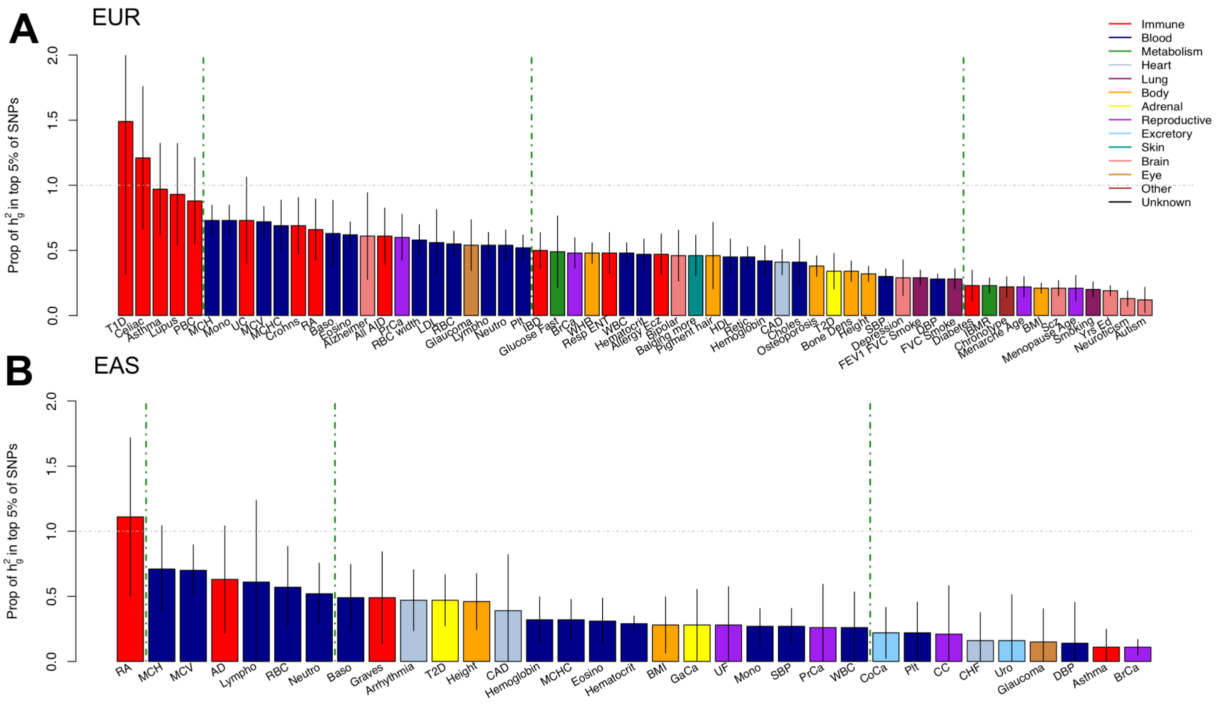


Figure B-5 legend. A) Common SNP heritability captured by the top 5% of SNPs according to the lead cell type association for each EUR GWAS. Lead association determined by largest τ^* estimate that is significantly positive. B) Similar for each EAS GWAS. Gray bars indicate the standard error of the heritability estimate. Color represents the category of the complex trait or disease.

Figure B-6. Comparison of heritability captured by lead IMPACT annotation vs lead cell-type-specific histone mark annotation.

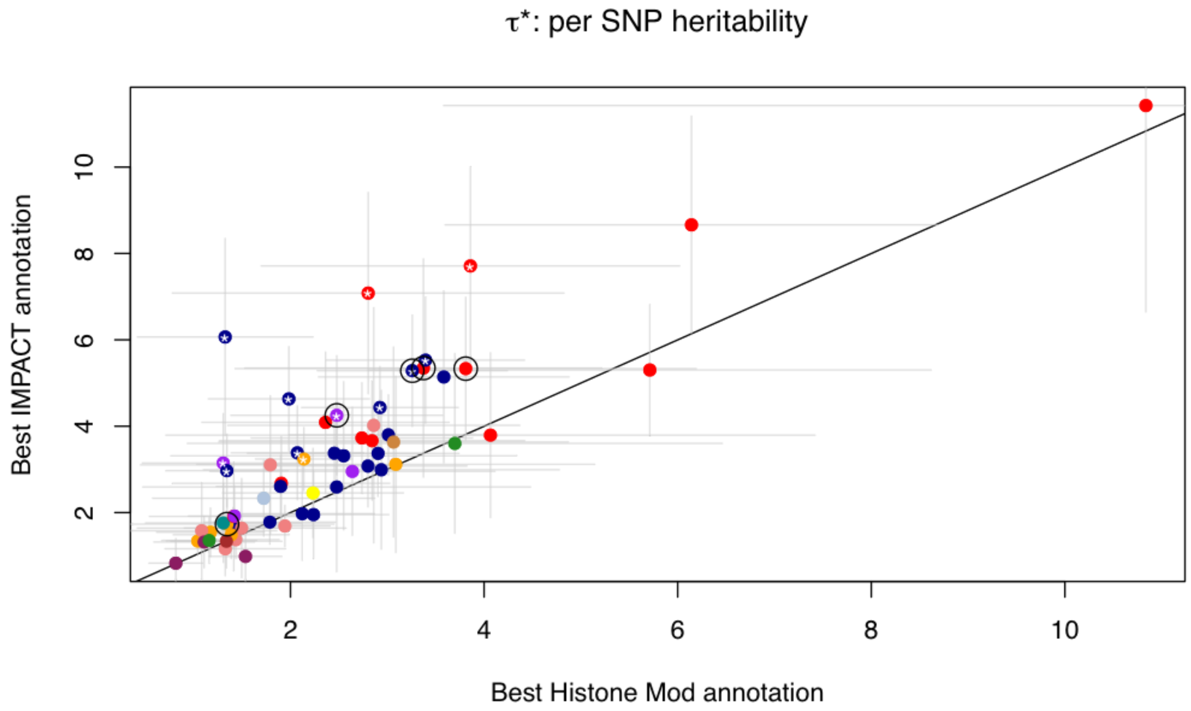


Figure B-6 legend. Comparison of two different functional annotations, IMPACT and cell type specific histone marks, to capture polygenic heritability assessed by quantifying * per-SNP heritability value. Circled are five representative traits used throughout the study: asthma, RA, PrCa, MCV, and height.

Figure B-7. Heritability captured by deep learning annotations.

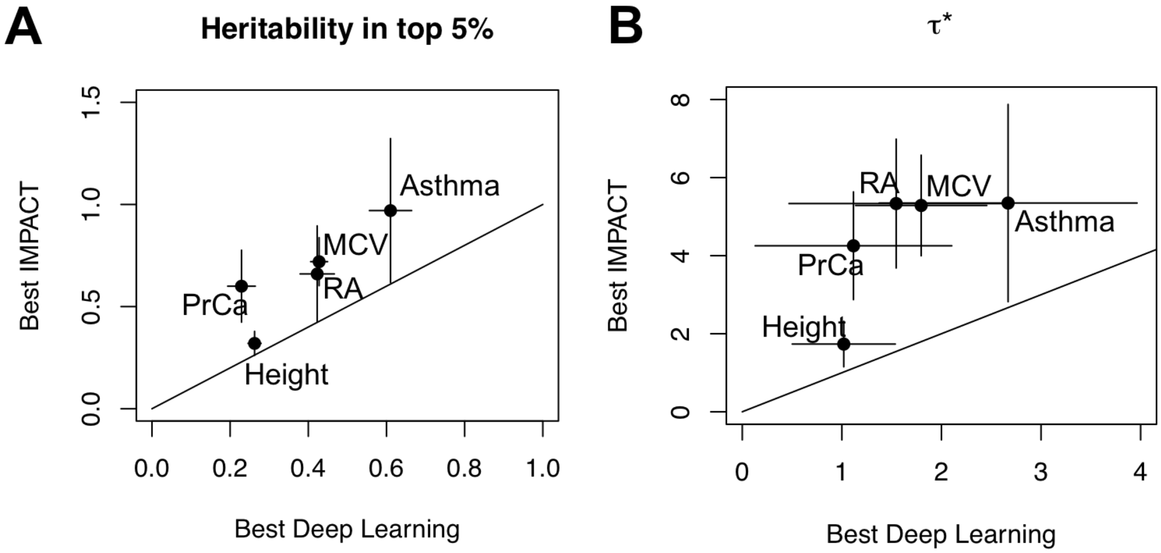


Figure B-7 legend. A) Among five representative traits, proportion of total SNP heritability captured by the lead IMPACT annotation compared to the lead deep learning annotation, from a set of 123 annotations. B) Among five representative traits, τ * of the lead IMPACT annotation compared to the lead deep learning annotation, from a set of 123 annotations.

Figure B-8. Heritability captured by deep learning annotations.

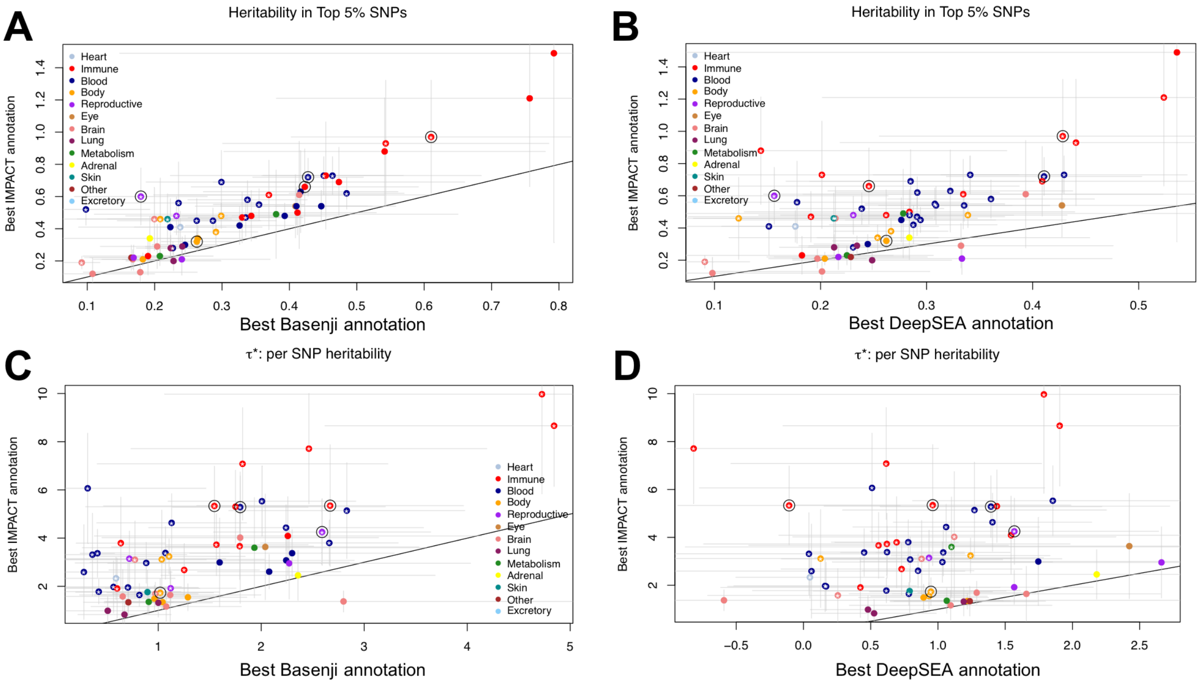


Figure B-8 legend. Proportion of total SNP heritability captured by top 5% of SNPs according to lead IMPACT annotation (y axis) and lead Basenji annotation (x axis) in panel A or lead DeepSEA annotation in panel B. Standardized annotation effect size τ^* according to lead IMPACT annotation (y axis) and lead Basenji annotation (x axis) in panel C or lead DeepSEA annotation in panel D.

Figure B-9. Conditional analyses to identify independent IMPACT associations.

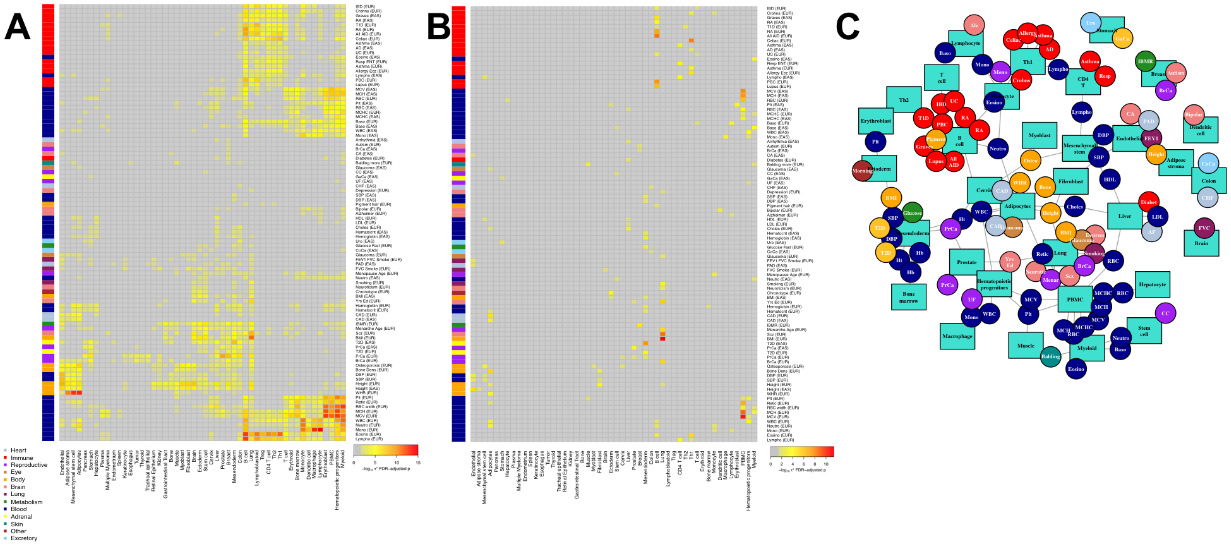


Figure B-9 legend. A) Stratification of IMPACT annotation associations by 50 cell types across the 95 polygenic traits and diseases of 111 with at least one association. For each cell type, the strongest annotation association is represented ($-log_{10}\ \tau * P$ value, FDR 5% adjusted). B) After four rounds of conditional analysis, non-independent associations were removed. Shown are the remaining independent annotation associations of the same 50 cell types and 95 traits. Color indicates $-log_{10}\ \tau * P$ value adjusted for FDR 5%; if more than one independent cell type association, $-log_{10}\ \tau *$ conditional P value adjusted for FDR 5% is indicated. C) Network of remaining independent associations, same information as in B), reveals clusters of regulatory modules that recapitulate known biology.

Figure B-10. Relation between heritability and number of independent regulatory annotations.

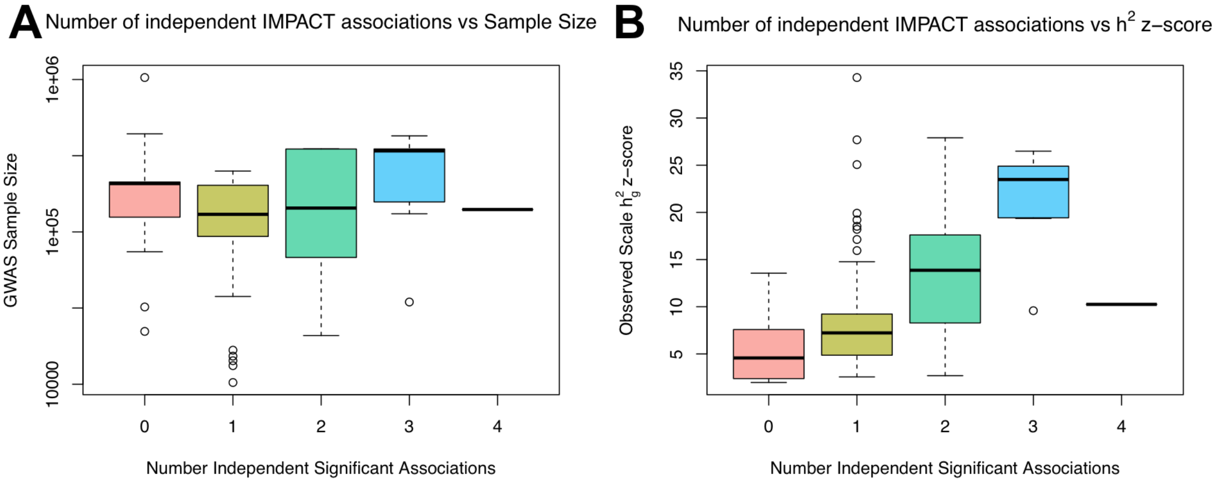


Figure B-10 legend. A) Number of independent IMPACT cell type associations is not significantly correlated with the sample size of the GWAS (*P* = 0.19). B) Number of independent associations is significantly positively correlated with the observed scale heritability z-score of the trait (*P* < 5.4e-9).

Figure B-11. Multi-ethnic regulatory concordance with IMPACT across traits.

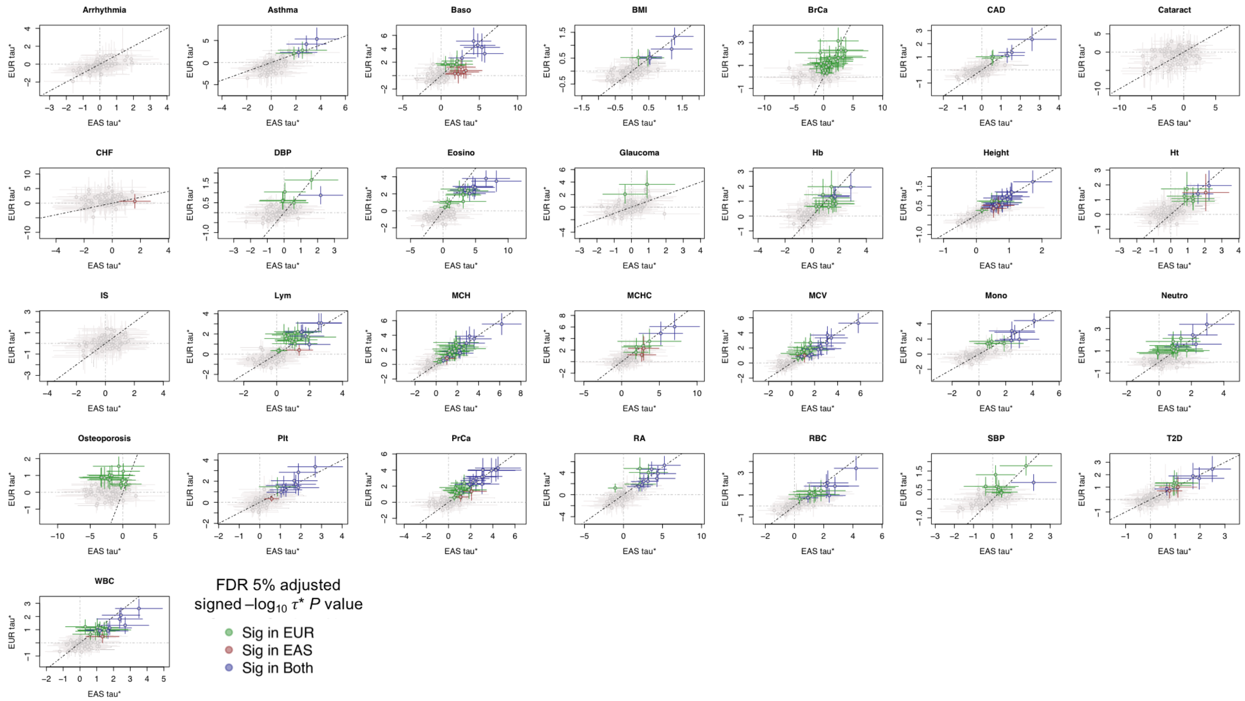


Figure B-11 legend. Common per-SNP heritability (τ *) estimate for sets of independent IMPACT cell type annotations across 29 traits. Dotted line is the identity line, y = x. τ * values with their standard errors are colored green if significantly positive in EUR and not EAS, red if significantly positive in EAS but not in EUR, green if significantly positive in both EUR and EAS, and gray if not significantly positive in either population.

Figure B-12. Multi-ethnic regulatory concordance with other functional annotations.

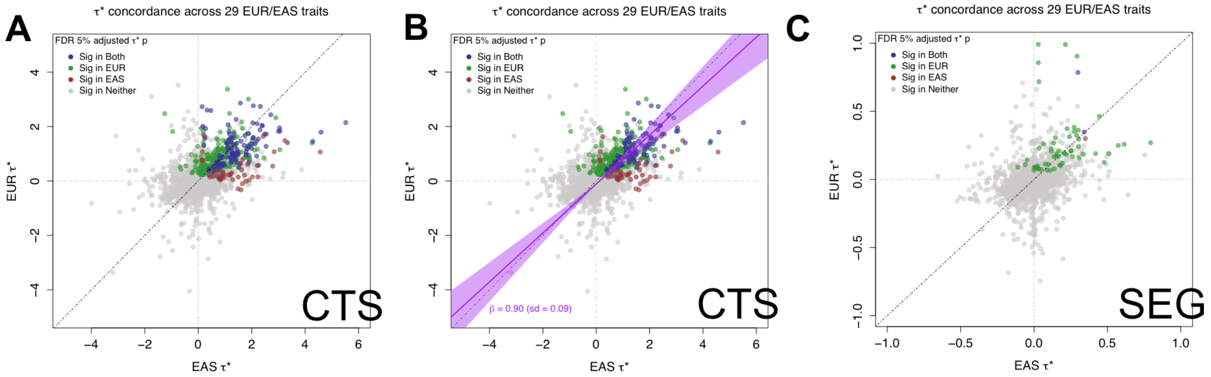


Figure B-12 legend. A) Common per-SNP heritability (τ *) estimate for sets of independent cell-type-specific histone mark annotations from Finucane et al Nature Genetics 2015 (EUR annotations) and Kanai et al Nature Genetics 2018 (EAS annotations) across 29 traits. B) As in A) after removing eight outlier annotations from “Sig in Both” category with noticeably larger EUR τ * and small EAS τ *, revealing a cross-ancestry relationship that is not dissimilar from identity. Line of best fit through annotations significant in both populations (dark purple line, 95% CI in light purple). C) As in A) for sets of independent cell-type-specifically expressed gene sets from Finucane et al Nature Genetics 2018 (EUR annotations) and Kanai et al Nature Genetics 2018 (EAS annotations). For all panels, the dotted line is the identity line, y = x. τ * values with their standard errors are colored green if significantly positive in EUR and not EAS, red if significantly positive in EAS but not in EUR, green if significantly positive in both EUR and EAS, and gray if not significantly positive in either population.

Figure B-13. Concordance of marginal effect sizes among selected SNPs using IMPACT across traits.

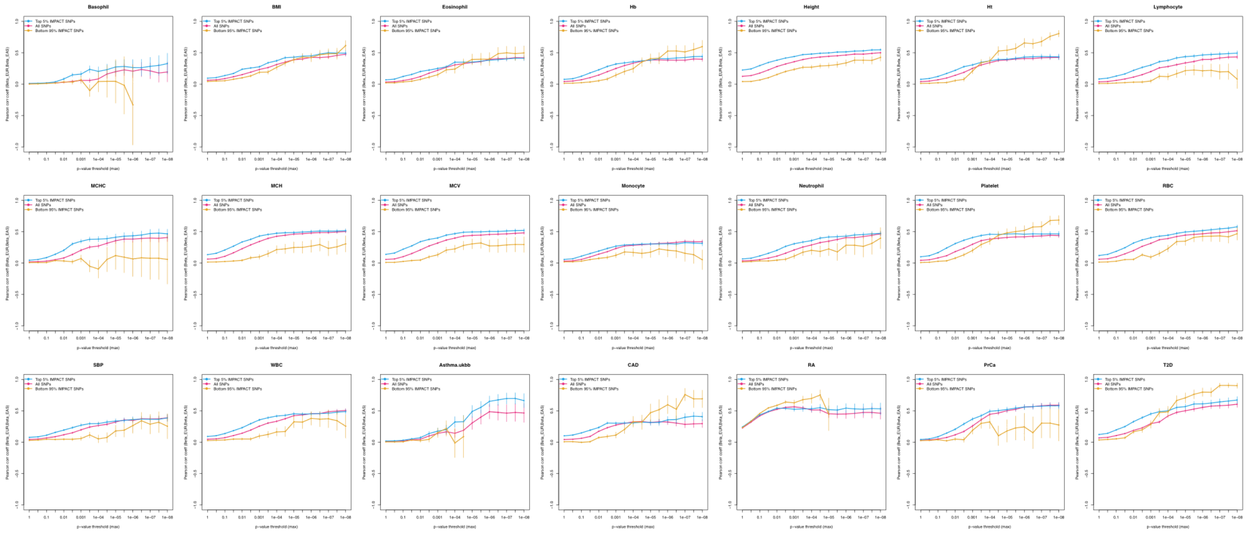


Figure B-13 legend. For 21 traits shared between EUR and EAS, effect size correlation (Pearson correlation coefficient) across 17 *P* value thresholds for three partitions of SNPs genome-wide: 1) lead SNPs with no IMPACT inference (red), 2) top 5% of SNPs according to the largest τ^* effect size IMPACT annotation (blue), and 3) the bottom 95% of SNPs according to the same IMPACT annotation (yellow). Vertical lines indicate one standard deviation of the correlation coefficient estimate.

Figure B-14. Concordance of marginal effect sizes among selected SNPs using other functional annotations.

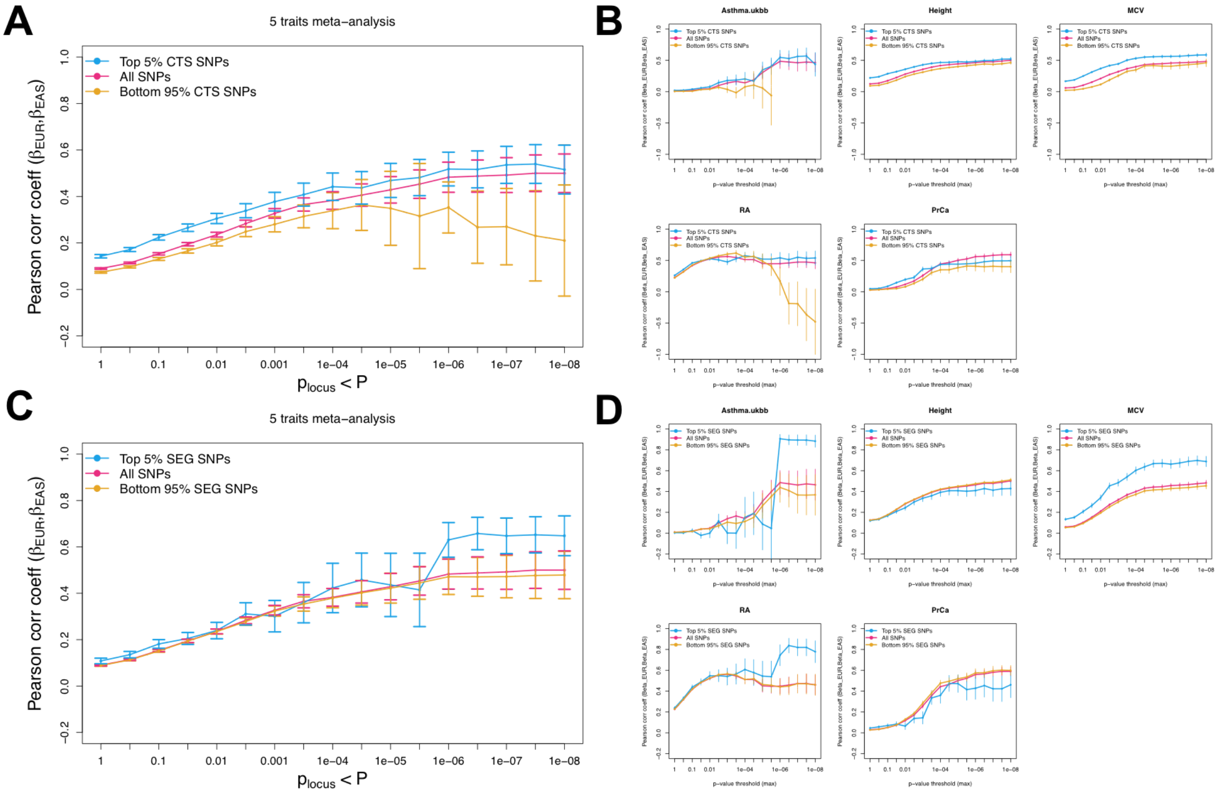


Figure B-14 legend. For 5 traits representing different biological underpinnings shared between EUR and EAS (subset of 21 investigated in our study), we report the effect size correlation (Pearson correlation coefficient) across 17 *P* value thresholds for three partitions of SNPs genome-wide: 1) lead SNPs with no functional inference (red), 2) top 5% of SNPs according to the largest τ * annotation effect size (blue), and 3) the bottom 95% of SNPs according to the same functional annotations (yellow). Here, we select the top annotation in two categories of previously published functional annotations: first, from LDSC-CTS annotations (meta-analysis in A, individual traits in B) and second, from LDSC-SEG annotations (meta-analysis in C, individual traits in D). Vertical lines indicate one standard deviation of the correlation coefficient estimate.

Figure B-15. Concordance of marginal effect sizes among selected SNPs using other functional annotations.

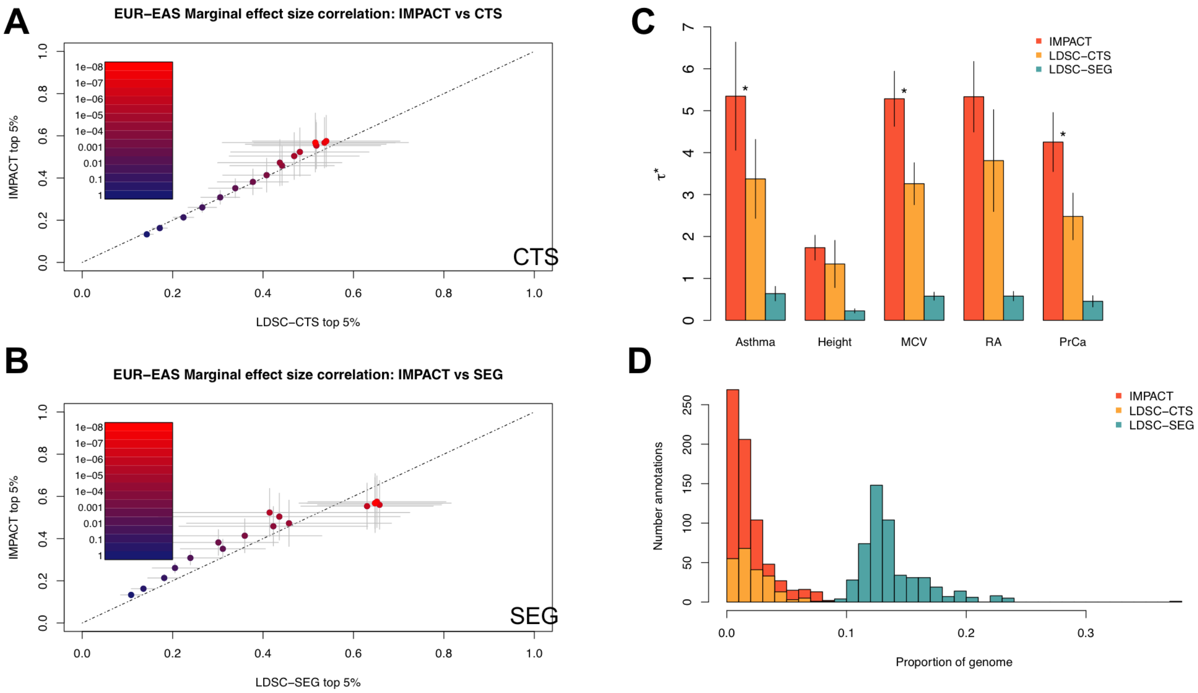


Figure B-15 legend. A) Comparison of top LDSC-CTS annotations in multi-ethnic effect size correlation analysis with top IMPACT annotations meta-analyzed over 5 traits. B) Similar to A) but for LDSC-SEG annotations C) τ * across the 5 selected traits reveals that IMPACT annotations are more strongly enriched for trait heritability than LDSC-CTS annotations (indicated by asterisk, difference of means *P* < 0.05) and consistently more than LDSC-SEG annotations. D) Distribution of annotation sizes for three different functional regimes: IMPACT (red), LDSC-CTS (yellow), LDSC-SEG (teal).

Figure B-16. Measuring concordance of heterozygosity among selected SNPs.

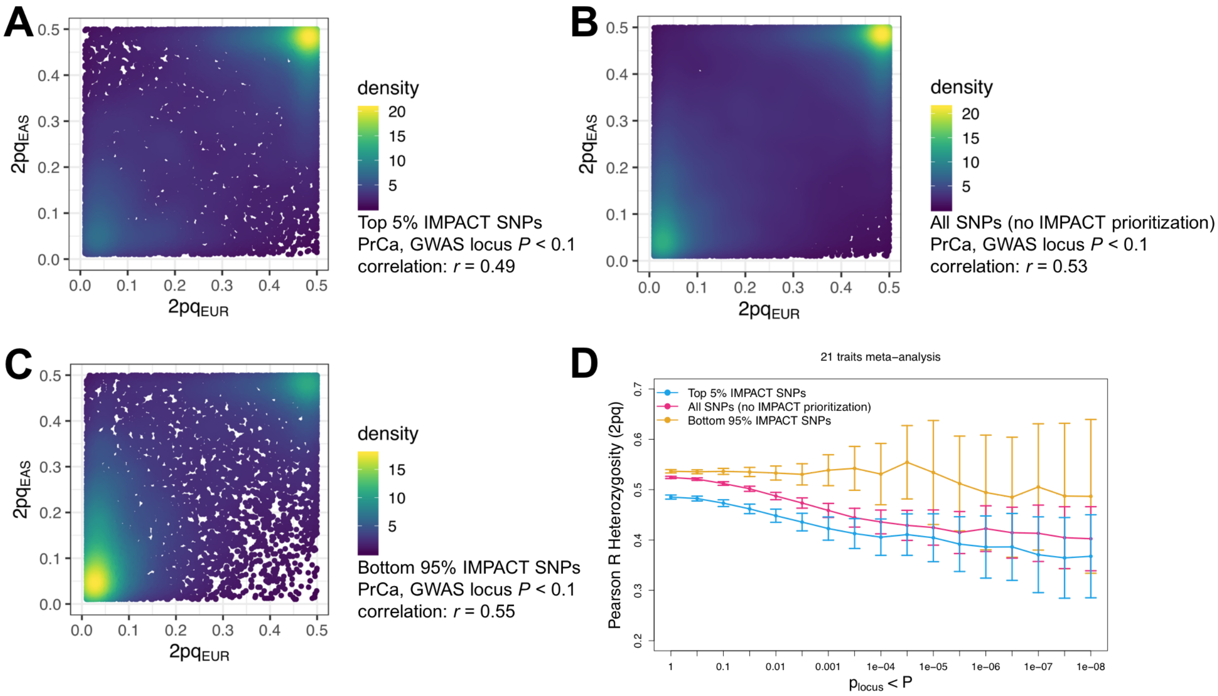


Figure B-16 legend. Population concordance of heterozygosity (2pq) among variants prioritized by IMPACT compared to standard P+T. A) Heterozygosity of variants from genome-wide EUR and EAS PrCa summary statistics in the top 5% of the lead IMPACT annotation for EUR PrCa. B) Heterozygosity of variants from genome-wide EUR and EAS PrCa summary statistics using standard P+T. C) Heterozygosity of variants from genome-wide EUR and EAS PrCa summary statistics in the bottom 95% of the lead IMPACT annotation for PrCa; mutually exclusive with SNPs in A). D) Meta-analysis of heterozygosity correlations between populations across 21 traits shared between EUR and EAS cohorts over 17 GWAS *P* value thresholds (with reference to the EUR GWAS).

Figure B-17. Measuring concordance of heterozygosity among selected SNPs across all traits.

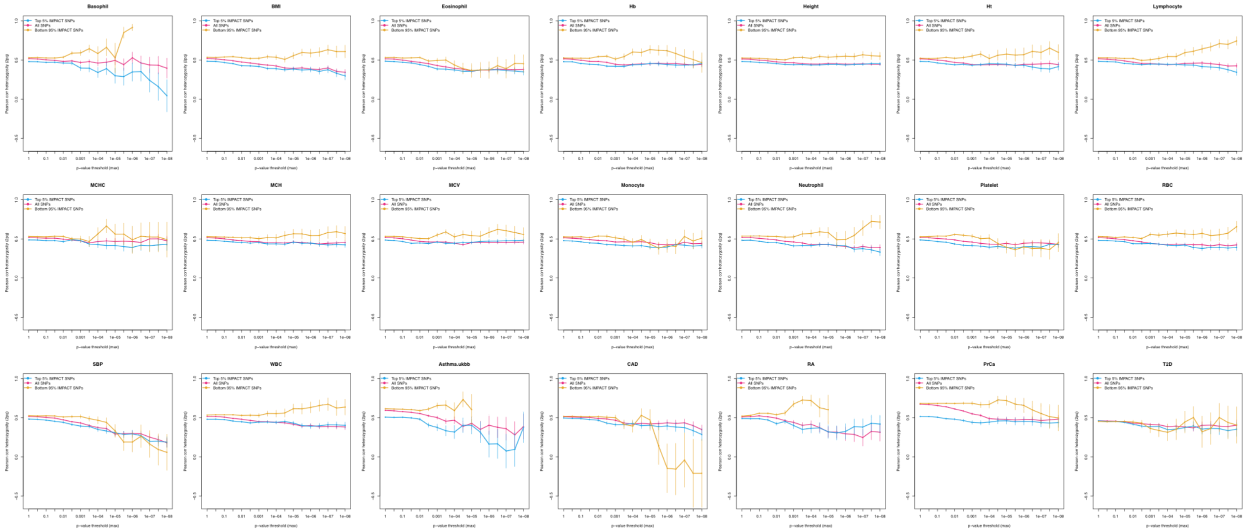


Figure B-17 legend. For 21 traits shared between EUR and EAS, heterozygosity (2pq) correlation (Pearson correlation coefficient) across 17 *P* value thresholds for three partitions of SNPs genome-wide: 1) lead SNPs with no IMPACT inference (red), 2) top 5% of SNPs according to the largest τ * effect size IMPACT annotation (blue), and 3) the bottom 95% of SNPs according to the same IMPACT annotation (yellow). Vertical lines indicate one standard deviation of the correlation coefficient estimate.

Figure B-18. Measuring the degree of population divergence among selected SNPs.

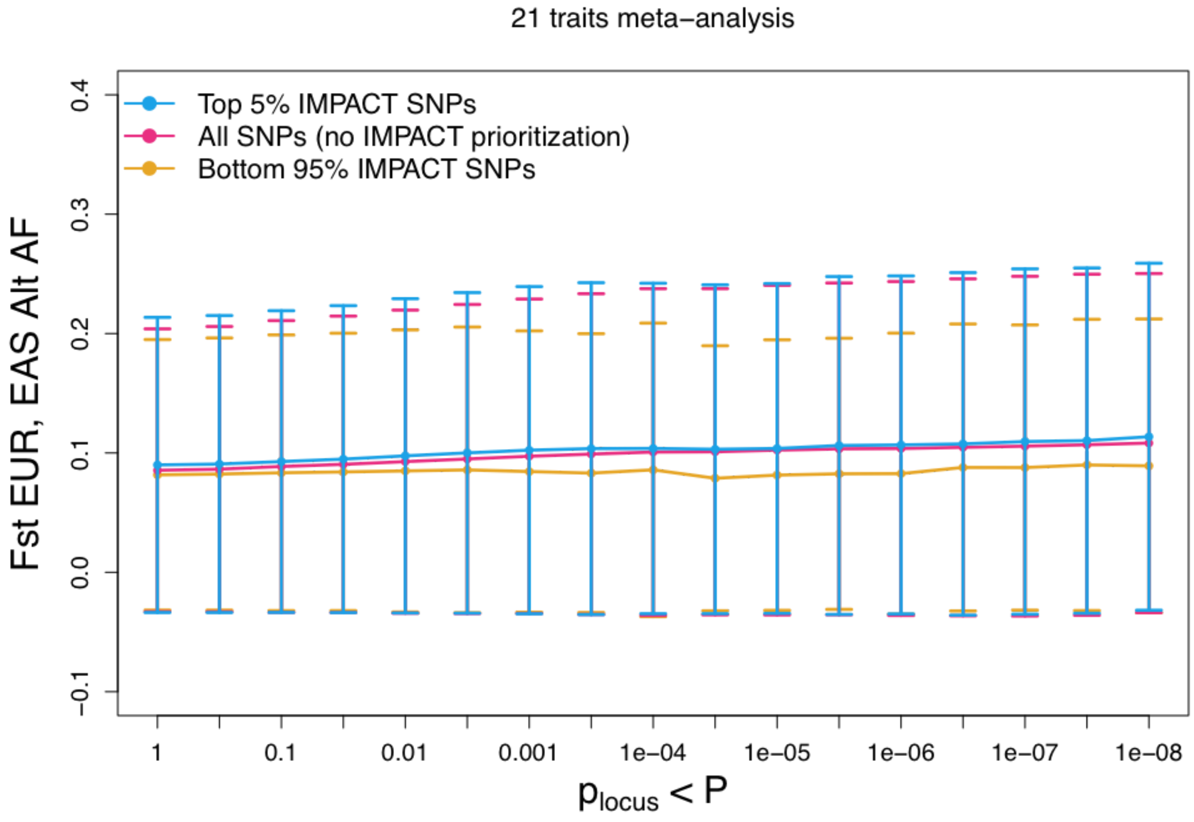


Figure B-18 legend. Population divergence, measured by F_{st}, where larger values indicate a reduction in heterozygosity, among variants prioritized by IMPACT compared to standard P+T. Meta-analysis of F_{st} between EUR and EAS populations across 21 traits shared between EUR and EAS cohorts over 17 GWAS *P* value thresholds (with reference to the EUR GWAS).

Figure B-19. Measuring the degree of population divergence among selected SNPs over all traits.

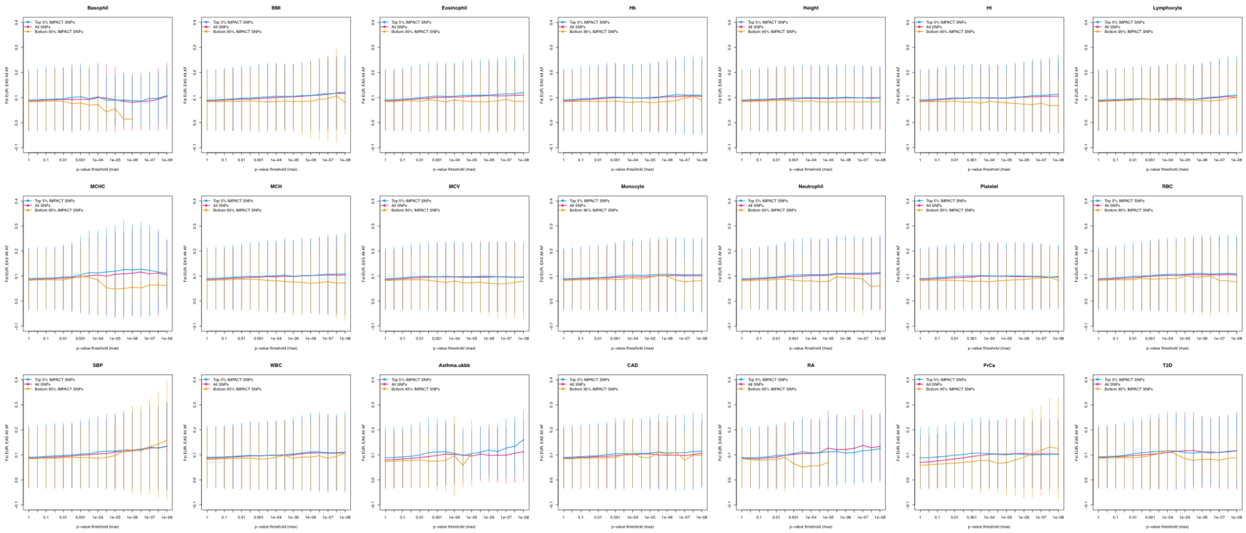


Figure B-19 legend. For 21 traits shared between EUR and EAS, we computed the average F_{st} , where large values indicate a reduction in heterozygosity, of sets of variants across 17 *P* value thresholds for three partitions of SNPs genome-wide: 1) lead SNPs with no IMPACT inference (red), 2) top 5% of SNPs according to the largest τ * effect size IMPACT annotation (blue), and 3) the bottom 95% of SNPs according to the same IMPACT annotation (yellow). Vertical lines indicate one standard deviation of the mean F_{st} estimate.

Figure B-20. PRS-EUR evaluated on all BBJ individuals across all traits.

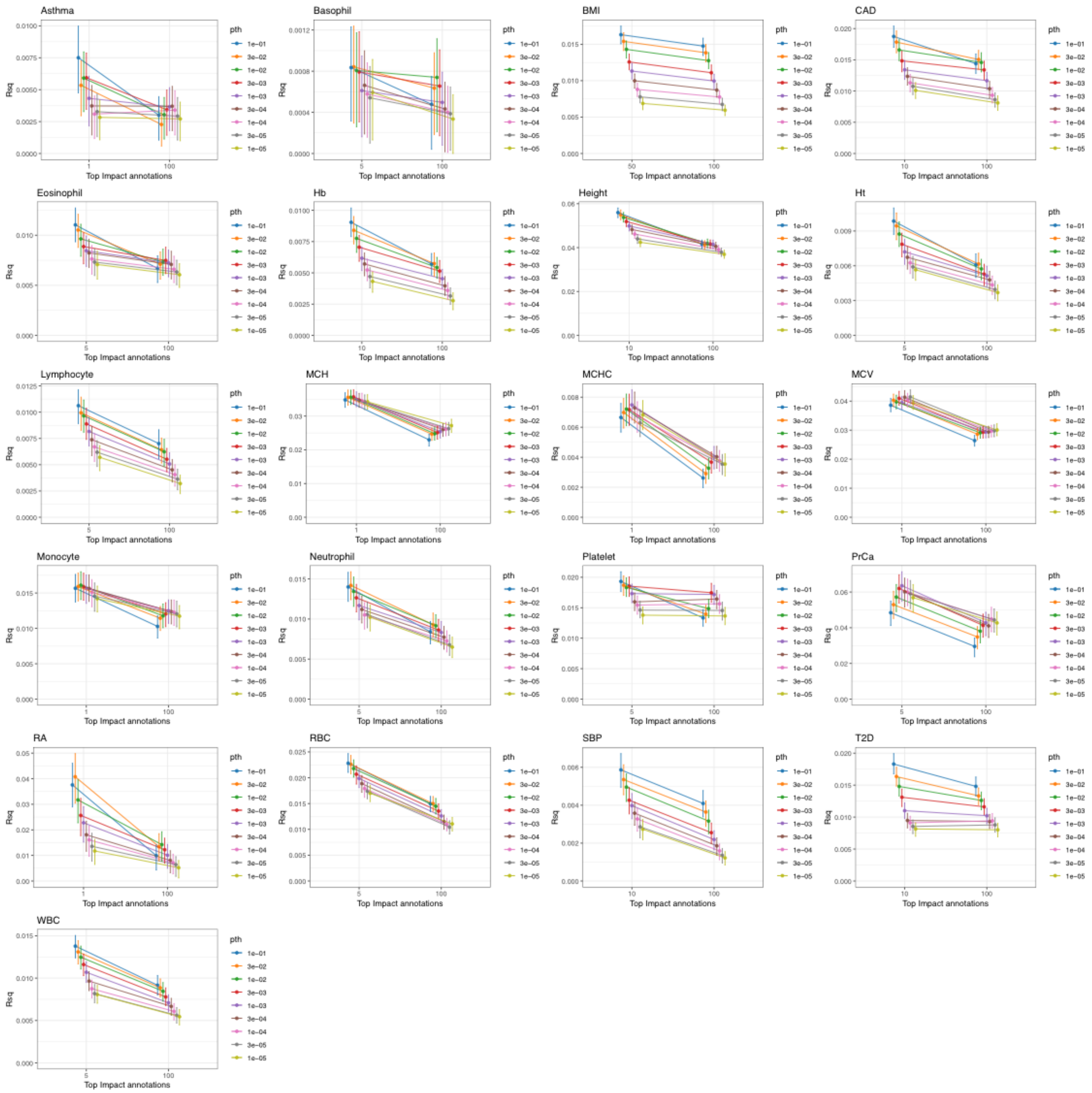


Figure B-20 legend. EUR PRS model evaluated on EAS individuals from BBJ. For each trait, we evaluate the predictive value of standard PRS models (top 100% of IMPACT SNPs) and functionally-informed PRS models (using a subset of SNPs prioritized by IMPACT). The top 100% of SNPs according to IMPACT represents the PRS model with no functional annotation information. Intervals represent the 95% confidence interval around the R^2 estimate. For quantitative traits, R^2 represents the proportion of variance captured by the linear PRS model.

For case control traits, R^2 represents the liability scale R^2 from the logistic regression PRS model.

Figure B-21. PRS-EUR evaluated on 5K BBJ individuals across all traits.

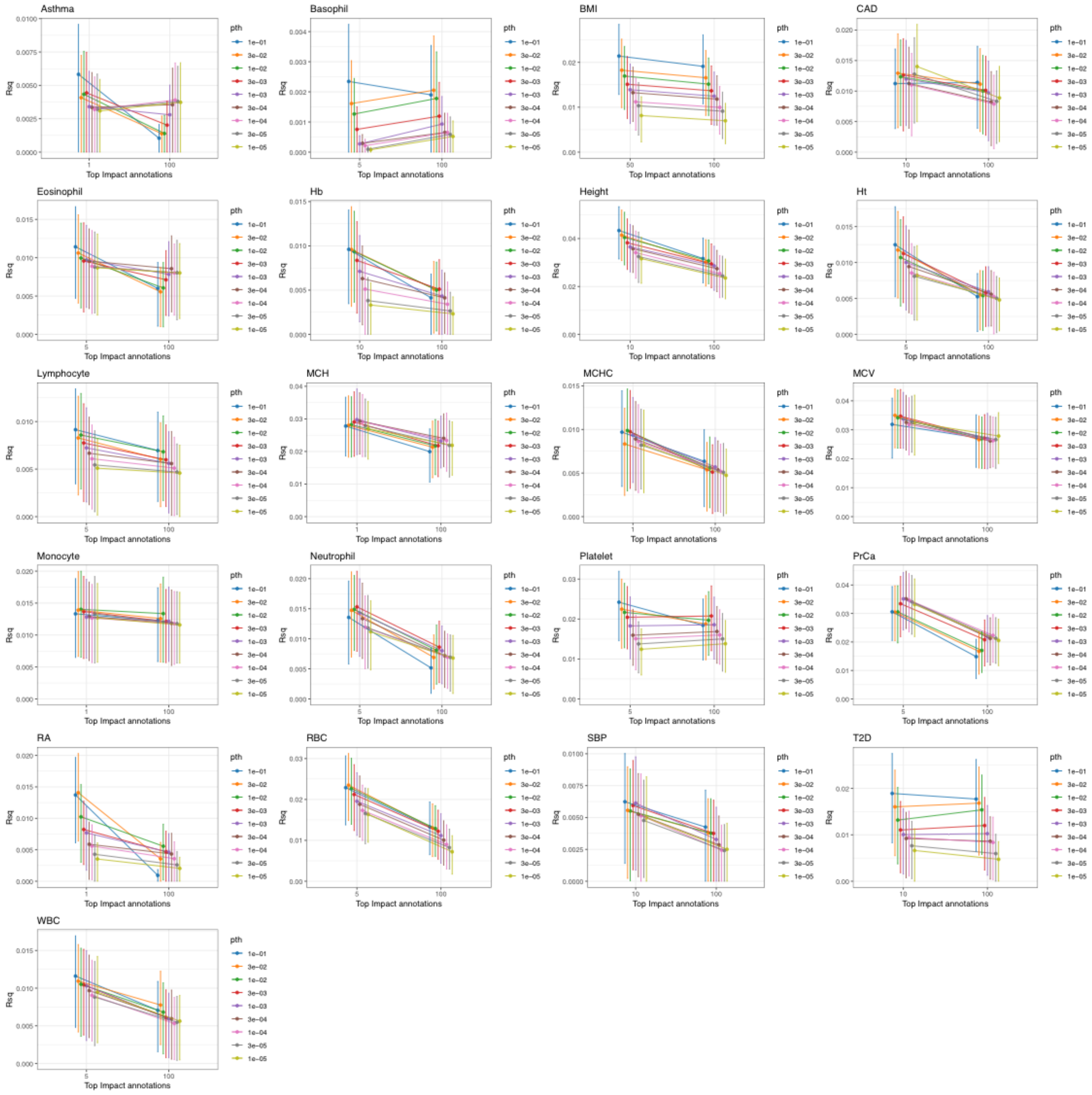


Figure B-21 legend. EUR PRS model evaluated on 5,000 randomly selected EAS individuals from BBJ. For each trait, we evaluate the predictive value of standard PRS models (top 100% of

IMPACT SNPs) and functionally-informed PRS models (using a subset of SNPs prioritized by IMPACT). Intervals represent the 95% confidence interval around the R^2 estimate. For quantitative traits, R^2 represents the proportion of variance captured by the linear PRS model. For case control traits, R^2 represents the liability scale R^2 from the logistic regression PRS model.

Figure B-22. PRS-EAS evaluated on 5K BBJ individuals across all traits.

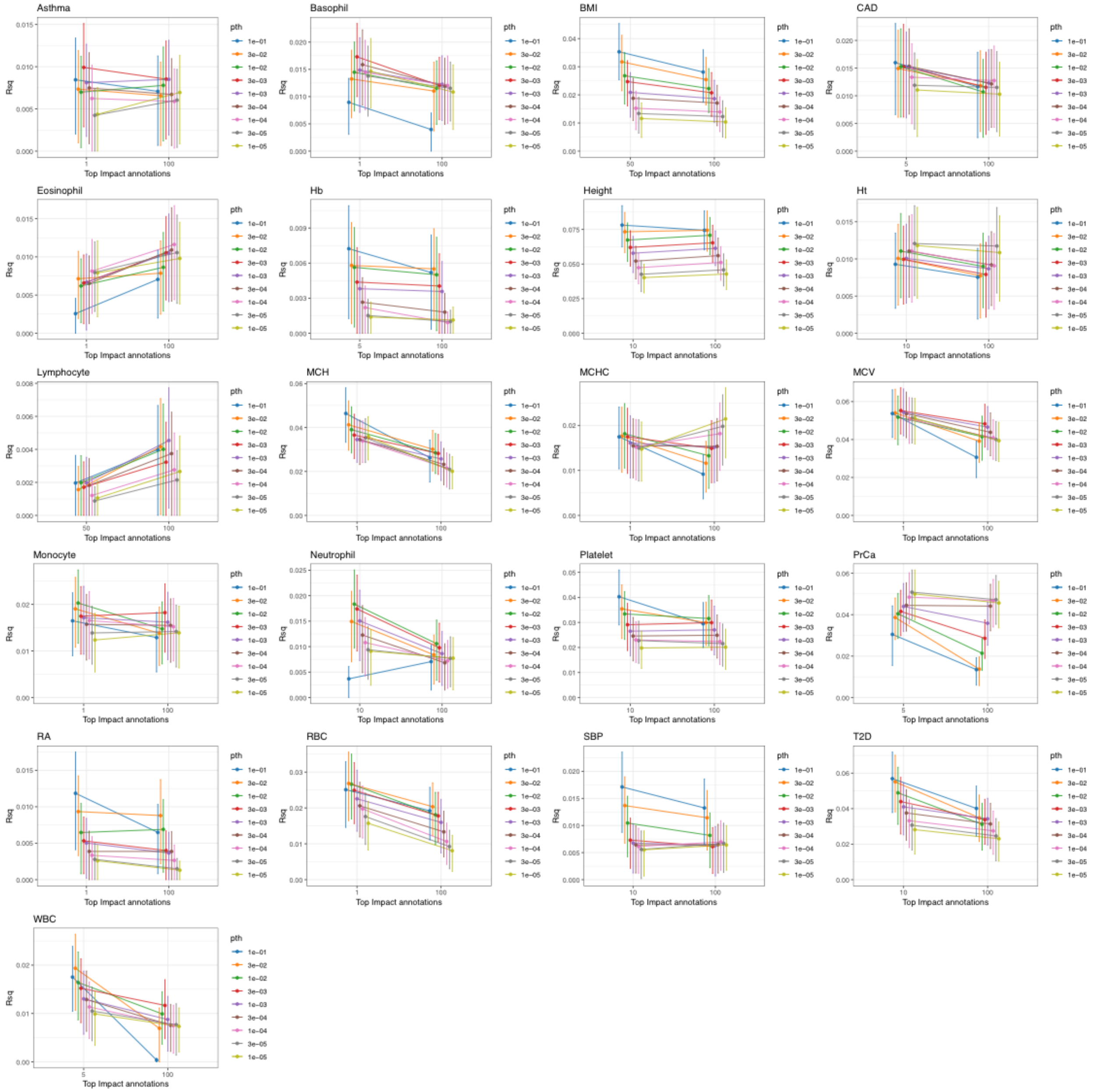


Figure B-22 legend. EAS PRS model evaluated on 5,000 non-overlapping EAS individuals from BBJ; these 5,000 individuals are the same as EAS test individuals in SF15. For each trait, we evaluate the predictive value of standard PRS models (top 100% of IMPACT SNPs) and functionally-informed PRS models (using a subset of SNPs prioritized by IMPACT). Intervals represent the 95% confidence interval around the R^2 estimate. For quantitative traits, R^2

represents the proportion of variance captured by the linear PRS model. For case control traits, R^2 represents the liability scale R^2 from the logistic regression PRS model.

Figure B-23. PRS evaluated in 5K BBJ individuals.

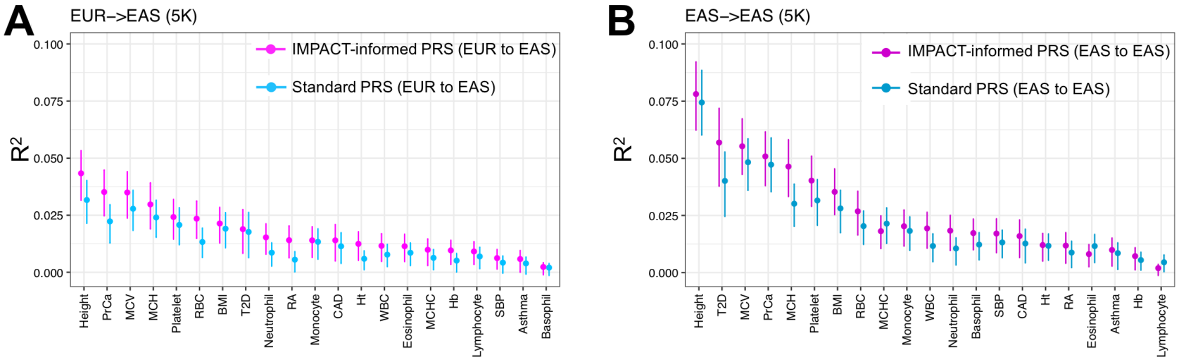


Figure B-23 legend. A) Phenotypic variance (R^2) in 5,000 BBJ individuals explained by IMPACT-informed PRS-EUR (dark pink) and standard PRS-EUR (dark blue). B) Phenotypic variance (R^2) in 5,000 BBJ individuals explained by IMPACT-informed PRS-EAS (light pink) and standard PRS-EAS (light blue). Error bars indicate 95% CI calculated via 1,000 bootstraps.

Figure B-24. Block jackknife of PRS estimates.

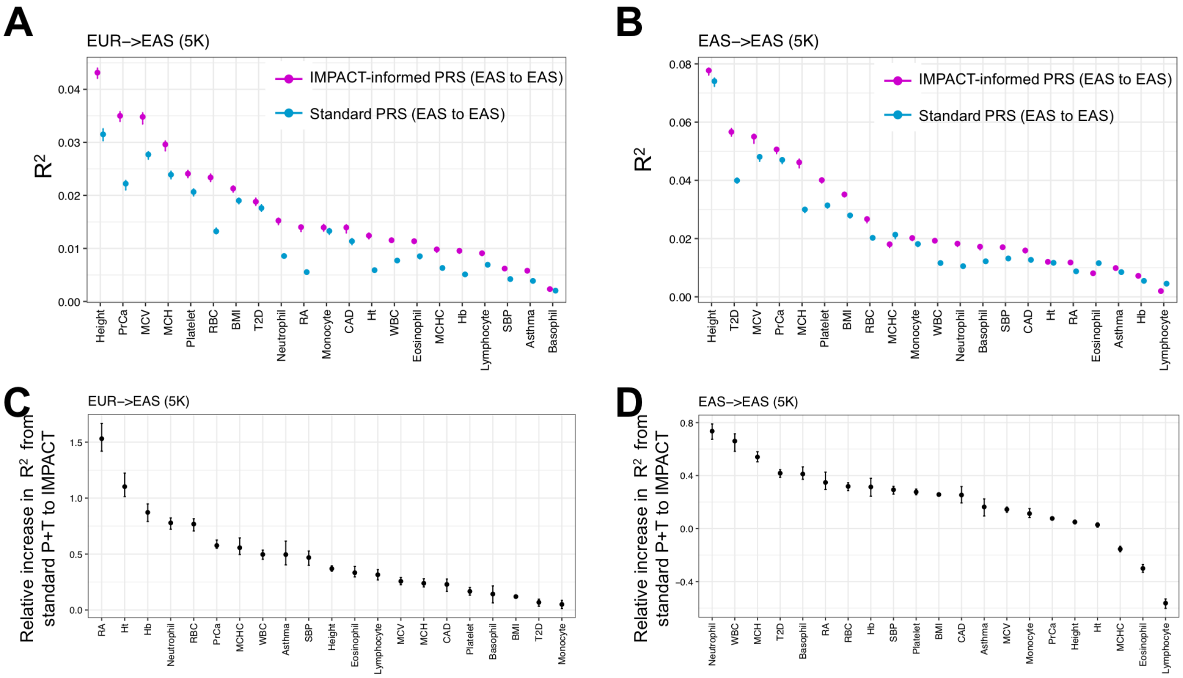


Figure B-24 legend. We recomputed confidence intervals around the R^2 estimates (panels A and B) and around the relative improvements in R^2 estimates of IMPACT PRS over standard P+T PRS (panels C and D) via block jackknife across the genome, using 200 adjacent equally-sized bins and iteratively removing variants within each bin and computing the R^2. A) Trans-ethnic analysis of EUR PRS to BBJ individuals. B) Within-population analysis of EAS PRS to BBJ individuals. Error bars indicate 95% CI around the R^2 estimates. C) Trans-ethnic analysis of EUR PRS to BBJ individuals, relative improvement in R^2 estimates defined as (IMPACT R^2 - standard P+T R^2) / standard P+T R^2. D) Within-population analysis of EAS PRS to BBJ individuals, relative improvement in R^2 estimates defined as (IMPACT R^2 - standard P+T R^2) / standard P+T R^2 .

Figure B-25. Liability versus logistic scale in PRS.

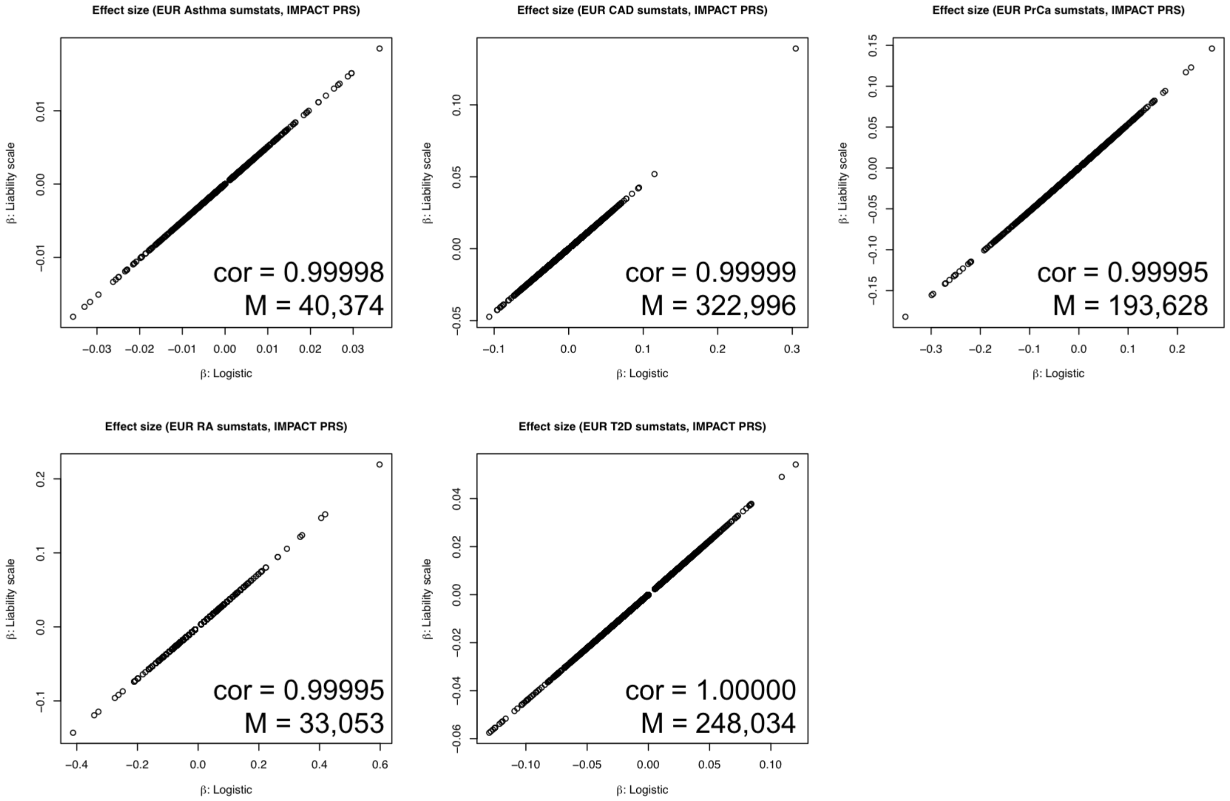


Figure B-25 legend. For each of five case/control diseases considered in PRS analyses, we computed the correlation of effect size estimates on the logistic scale versus the liability scale. The set of variants selected for each disease corresponds to the IMPACT-informed PRS model with the highest R^2.

Figure B-26. Liability versus logistic scale in PRS.

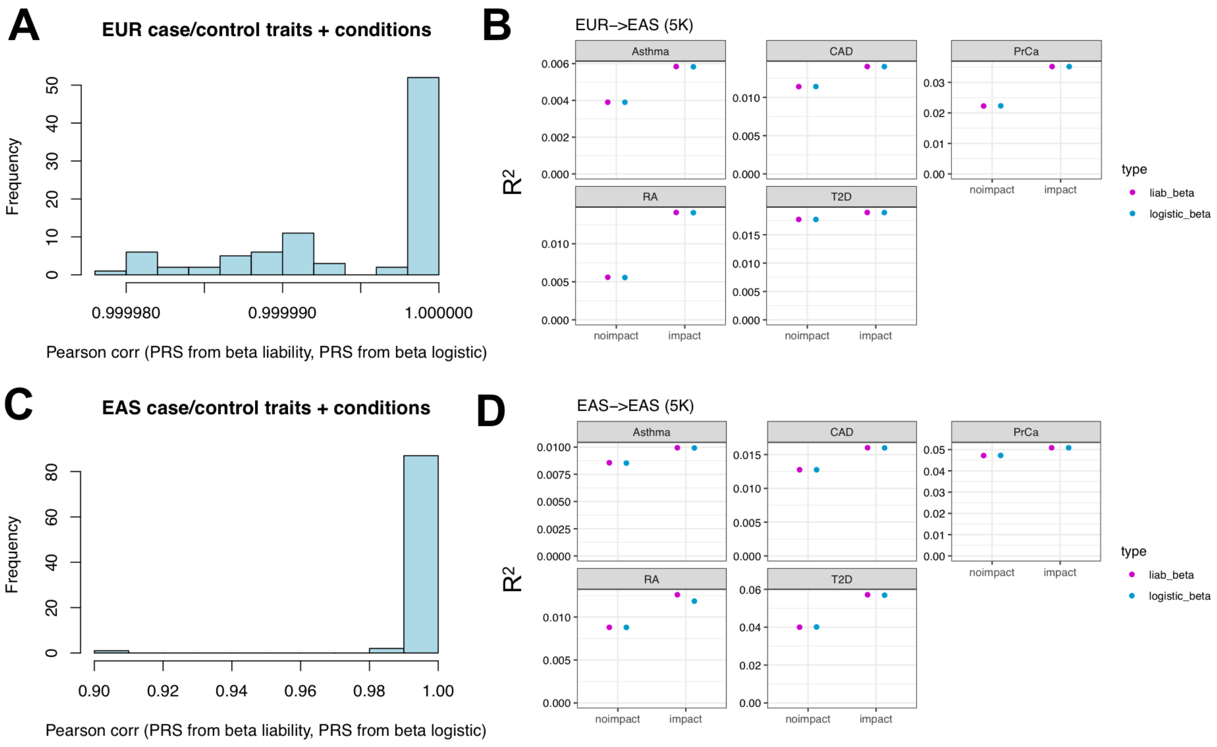


Figure B-26 legend. For each of five case/control diseases considered in PRS analyses, we computed the correlation of PRS values based on EUR effect size estimates calculated on the logistic scale versus the liability scale (panel A for PRS-EUR and panel C for PRS-EAS). All sets of variants were considered for this analysis, e.g. 9 *P* value thresholds x 2 model types (IMPACT/standard PRS) x 5 case/control diseases = 90. We also compare logistic and liability scale PRS R^2 between IMPACT-informed and standard P+T models (panel B for PRS-EUR and panel D for PRS-EAS). For this analysis, we only considered the *P* value threshold that achieved the highest R^2 for IMPACT and standard P+T models.

Figure B-27. IMPACT stabilization of PRS distributions using EUR GWAS data.

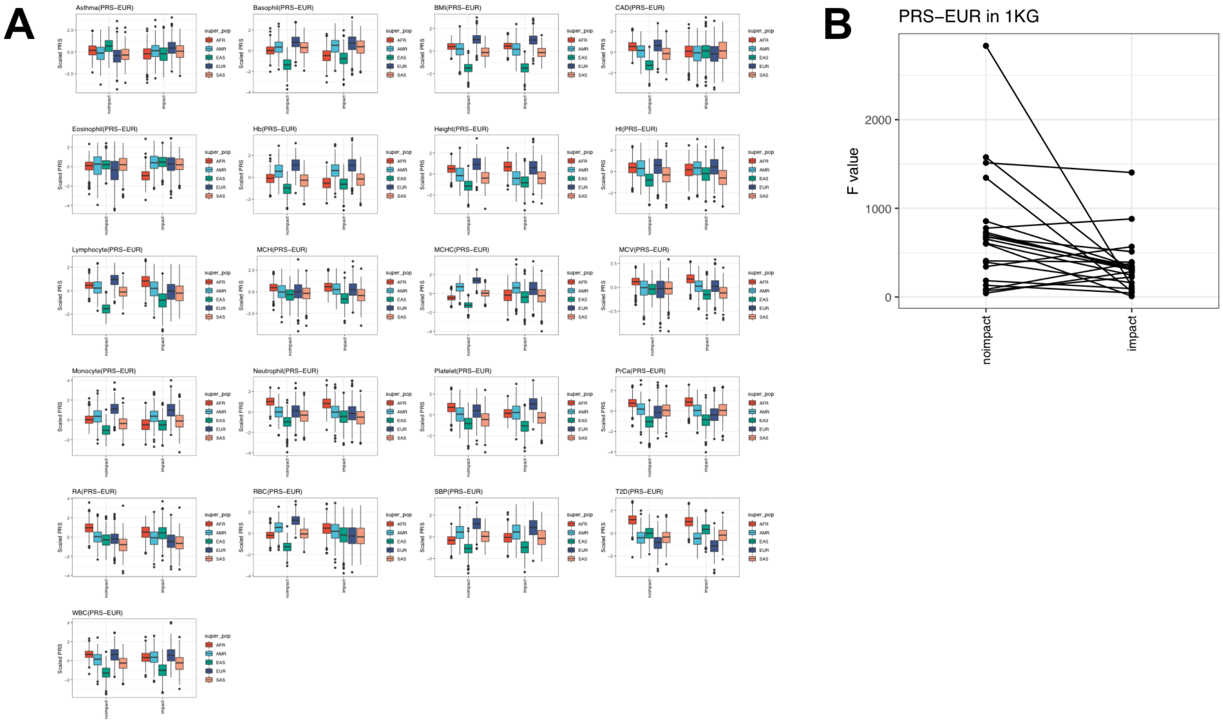


Figure B-27 legend. A) For each of 21 traits considered in the EUR PRS analyses, we compare the variance in the polygenic risk scores based on standard P+T and IMPACT-informed P+T using the model that achieved the highest R^2. B) We used anova to compare the observed variance of PRS distributions across the five different 1000G populations, for each trait between standard P+T PRS and IMPACT-informed PRS, by computing F-statistics.

Figure B-28. IMPACT stabilization of PRS distributions using EAS GWAS data.

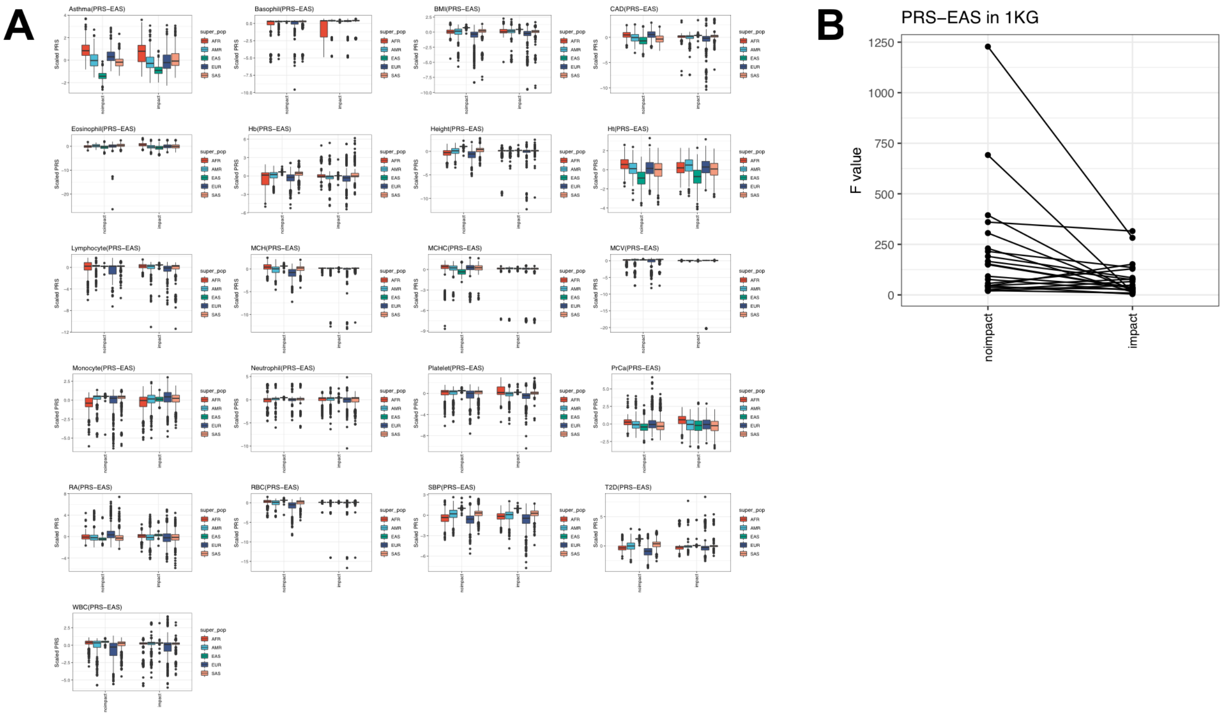
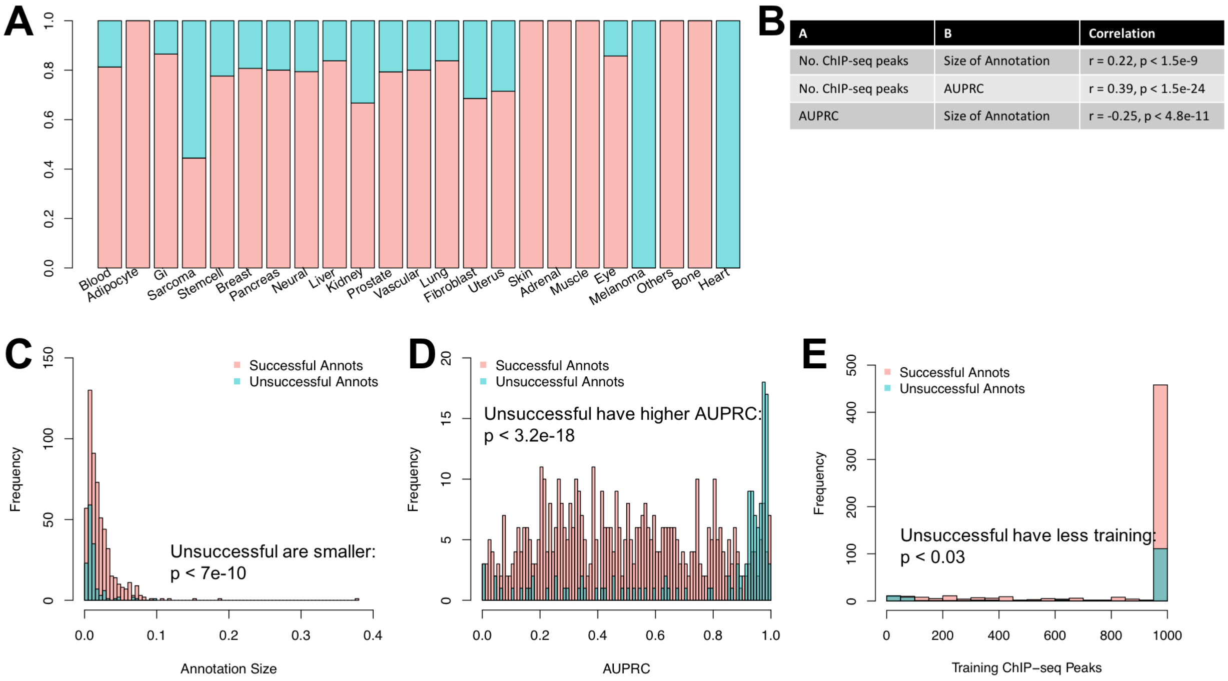


Figure B-28 legend. A) For each of 21 traits considered in the EAS PRS analyses, we compare the variance in the polygenic risk scores based on standard P+T and IMPACT-informed P+T using the model that achieved the highest R^2 . B) We used anova to compare the observed variance of PRS distributions across the five different 1000G populations, for each trait between standard P+T PRS and IMPACT-informed PRS, by computing F-statistics.

Figure B-29. Summary of 707 IMPACT annotations across cell types.



A	B	Correlation
No. ChIP-seq peaks	Size of Annotation	r = 0.22, p < 1.5e-9
No. ChIP-seq peaks	AUPRC	r = 0.39, p < 1.5e-24
AUPRC	Size of Annotation	r = -0.25, p < 4.8e-11

Figure B-29 legend. A) Distribution of annotation size (average IMPACT score over annotated SNPs) for “successful” and “unsuccessful” annotations. B) Distribution of TF binding model AUPRC for “successful” and “unsuccessful” annotations. C) Distribution of training set size (number of TF ChIP-seq peaks) for “successful” and “unsuccessful” annotations. D) Correlation of metadata factors of IMPACT annotations: number of ChIP-seq peaks available to training data, AUPRC of TF binding prediction model, and annotation size. E) For each tissue type category of IMPACT annotation, the proportion of annotations that were significantly associated with at least one polygenic trait or disease (“successful”) is indicated by the height of the pink bar. “Unsuccessful” annotations were not found to be significantly associated with any phenotype and are indicated by the green bar. For example, heart-labeled annotations had no significant associations.

Figure B-30. Pairwise trait regulatory and genetic correlation.

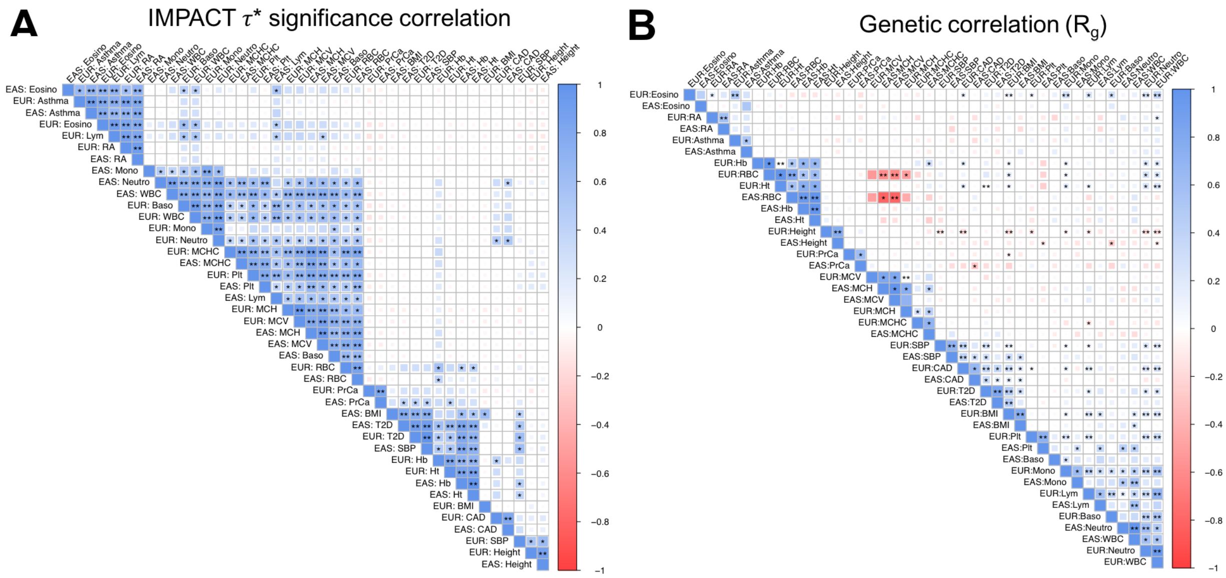


Figure B-30 legend. A) Pairwise correlation of IMPACT functional annotations' τ * significance across 42 traits, accounting for 21 unique phenotypes (those with at least one significant IMPACT association in both EUR and EAS) and two populations. * indicates FDR-adjusted *P* < 0.05, ** indicates FDR-adjusted *P* < 1e-10. B) Pairwise genetic correlation across the same 42 traits as in (A). * indicates nominal *P* < 0.05, ** indicates nominal *P* < 1e-10.

Figure B-31. Cell-type-combined annotations improve heritability capture.

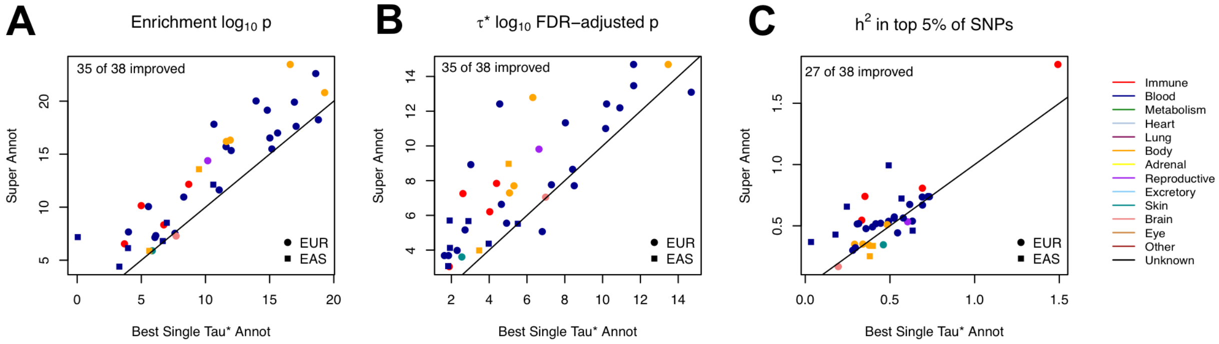


Figure B-31 legend. Comparison of heritability metrics between the lead annotation and the composite annotation, created from independently associated IMPACT annotations. A) Statistical significance of the enrichment estimate. B) Statistical significance of the τ * S-LDSC regression coefficient estimate. C) Proportion of observed scaled heritability in the top 5% SNPs scored by IMPACT.

Figure B-32. IMPACT selects variants with more concordant effect size estimates especially at lenient *P* values.

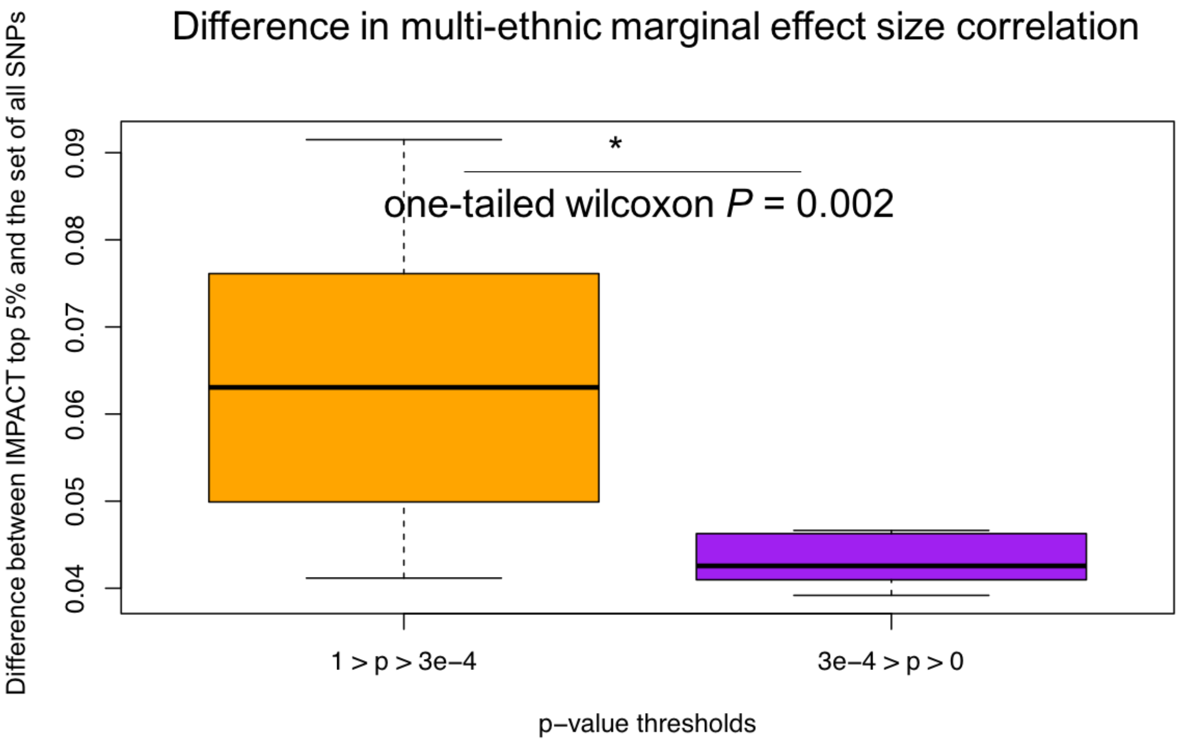
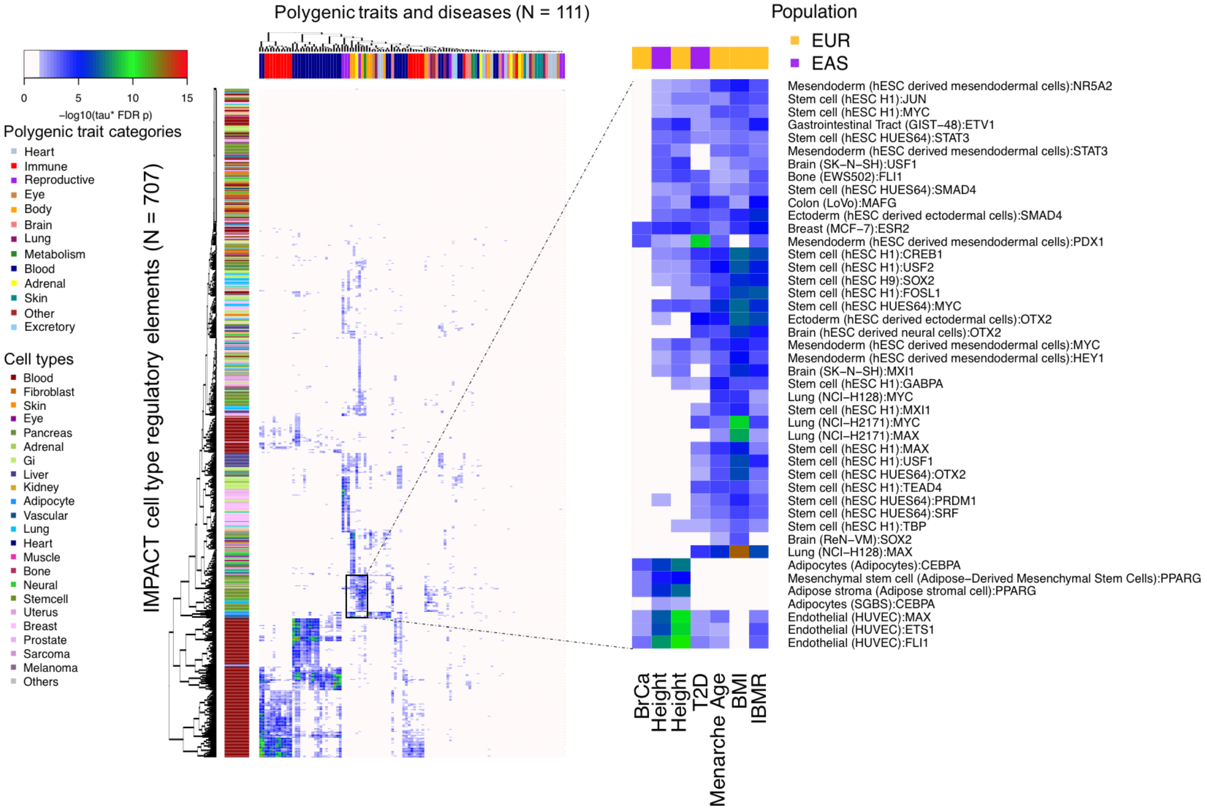


Figure B-32 legend. Improvement by functional data (IMPACT top 5% SNP selection) varies by *P* value threshold. Improvement is greatest when p-values are lenient (orange). Improvement is minimized when the EUR GWAS *P* value is near or past the genome-wide significant threshold (purple).

Ext. Data B-1. Overview of 707 by 111 possible annotation-trait associations.



Ext. Data B-1 legend. Significant cell type-phenotype associations across 707 IMPACT regulatory annotations and 111 complex traits and diseases at τ * 5% FDR, color indicates -log10 FDR 5% adjusted *P* value of τ *. Zooms shows particular cell type categories enriched for polygenic trait associations.

Appendix C

Supplementary Information for Chapter 4

Supplementary Figures

Figure C-1.

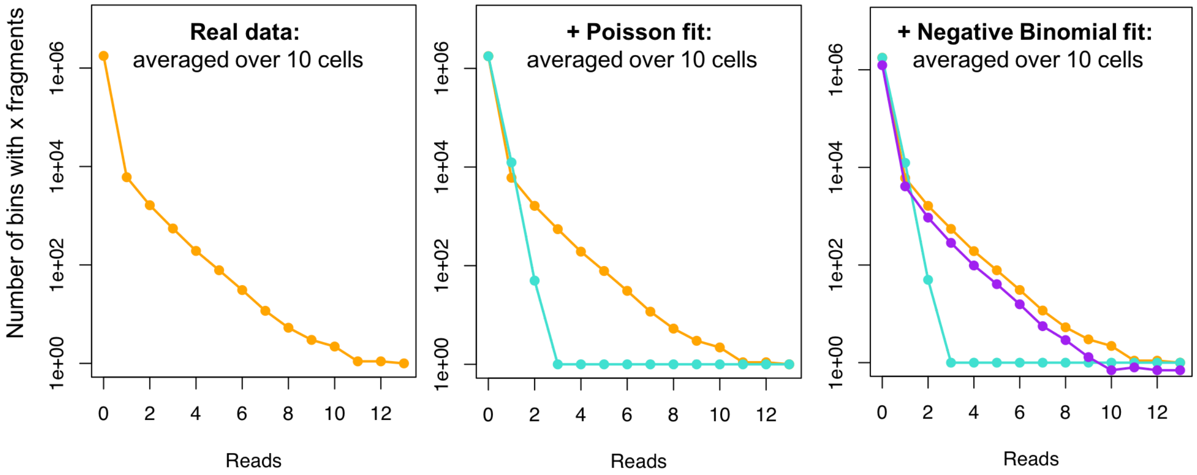


Figure C-1 legend. scATACseq fragment count data follows a negtaive binomial distribution.

Supplementary Tables

Table C-1. 37 IMPACT tracks use for cell-specific regulatory inference.

Cell type	TF	Cell type	TF
Treg	FOXP3	Breast (ZR-75-1)	HSF1
T cell (CCRF)	GATA3	Breast (MCF-7)	RXRA
T cell (T-ALL)	GATA3	Prostate (C4-2)	FOXA1
Treg	RUNX1	Prostate (NCI-H660)	ESR1
B cell	PAX5	Myoblast	MYOD1
B cell (OCILY10)	STAT3	Myotube	MYOD1

B cell (OCILY3)	STAT3	Stem cell (hESC H9)	SMAD4
B cell (GM10847)	CTCF	Stem cell (hESC H9)	SMAD3
Myeloid (K562)	CUX1	Liver (HepG2)	MAFK
PBMC	IRF5	Colon (HT-29)	HSF1
Monocyte	CEBPB	Colon (LoVo)	E2F3
Macrophage	CEBPB	Colon (LoVo)	SOX9
Monocyte (THP-1)	PPARG	Colon (LoVo)	BCL6
Myeloid (K562)	CEBPA	Colon (HCT 116)	FOSL1
Lung (NCI-H2171)	MYC	Plasma	MAX
Lung (NCI-H1703)	HSF1	Mesendoderm	EOMES
Lung (HCC95)	SOX2	Ectoderm	PAX6
Lung (A549)	SMAD3	Adipocytes (SGBS)	PPARG
Breast (MCF-7)	GATA3		

Appendix D

Bibliography

1. Amariuta, T., Luo, Y., Knevel, R., Okada, Y. & Raychaudhuri, S. Advances in genetics toward identifying pathogenic cell states of rheumatoid arthritis. *Immunol. Rev.* (2019). doi:10.1111/imr.12827
2. Stastny, P. Mixed lymphocyte cultures in rheumatoid arthritis. *J. Clin. Invest.* **57**, 1148 (1976).
3. Cornelis, F. *et al.* New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proceedings of the National Academy of Sciences* **95**, 10746–10750 (1998).
4. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228–237 (2003).
5. Begovich, A. B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet.* **75**, 330–337 (2004).
6. Suzuki, A. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34**, 395–402 (2003).
7. Plenge, R. M. *et al.* Replication of putative candidate-gene associations with rheumatoid arthritis in> 4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am. J. Hum. Genet.* **77**, 1044–1060 (2005).
8. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Kochi, Y. *et al.* A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* **42**, 515–519 (2010).
10. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).

11. Hu, X. *et al.* Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells. *PLoS Genet.* **10**, e1004404 (2014).

12. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).

13. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. *Am. J. Hum. Genet.* **89**, 496–506 (2011).

14. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

15. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

16. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2014).

17. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).

18. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

19. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).

20. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

21. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).

22. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

23. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

24. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

25. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

26. Reshef, Y. A. *et al.* Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).

27. Amariuta, T. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *Am. J. Hum. Genet.* **104**, 879–895 (2019).

28. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

29. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

30. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

31. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

32. Consortium, T. E. P. & The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

33. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

34. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

35. Chen, X., Yu, B., Carriero, N., Silva, C. & Bonneau, R. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res.* **45**, 4315–4329 (2017).

36. Karimzadeh, M. & Hoffman, M. M. Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome. *bioRxiv* 168419 (2018).

37. Soderquest, K. *et al.* Genetic variants alter T-bet binding and gene expression in mucosal inflammatory disease. *PLoS Genet.* **13**, e1006587 (2017).

38. Hertweck, A. *et al.* T-bet Activates Th1 Genes through Mediator and the Super Elongation Complex. *Cell Rep.* **15**, 2756–2770 (2016).

39. Gustafsson, M. *et al.* A validated gene regulatory network and GWAS identifies early regulators of T cell–associated diseases. *Sci. Transl. Med.* **7**, 313ra178–313ra178 (2015).

40. Tripathi, S. K. *et al.* Genome-wide Analysis of STAT3-Mediated Transcription during Early Human Th17 Cell Differentiation. *Cell Rep.* **19**, 1888–1901 (2017).

41. Schmidl, C. *et al.* The enhancer and promoter landscape of human regulatory and conventional T-cell subpopulations. *Blood* **123**, e68–e78 (2014).

42. Wang, C. *et al.* Genome-wide profiling of target genes for the systemic lupus erythematosus-associated transcription factors IRF5 and STAT4. *Ann. Rheum. Dis.* **72**, 96–103 (2013).

43. Qiao, Y. *et al.* Synergistic Activation of Inflammatory Cytokine Genes by Interferon-γ-Induced Chromatin Remodeling and Toll-like Receptor Signaling. *Immunity* **39**, 454–469 (2013).

44. Pham, T.-H. *et al.* Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood* **119**, e161–e171 (2012).

45. Dimitrova, L. *et al.* PAX5 overexpression is not enough to reestablish the mature B-cell phenotype in classical Hodgkin lymphoma. *Leukemia* **28**, 213–216 (2014).

46. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol. Cell* **52**, 25–36 (2013).

47. Kasowski, M. *et al.* Variation in Transcription Factor Binding Among Humans. *Science* **328**, 232–235 (2010).

48. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

49. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).

50. Ishigaki, K. *et al.* Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* **49**, 1120–1125 (2017).

51. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*

506, 376–381 (2014).

52. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics* **50**, 390–400 (2018).
53. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* 1 (2018).
54. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
55. Consortium, T. 1000 G. P. & The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
56. Westra, H.-J. *et al.* Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet.* **50**, 1366–1374 (2018).
57. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics* **44**, 1336–1340 (2012).
58. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
59. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
60. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21**, 1769–1776 (2011).
61. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 6131–6138 (2014).
62. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22**, 1723–1734 (2012).
63. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
64. Davenport, E. E. *et al.* Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial.

Genome Biol. **19**, 168 (2018).

65. Liu, X. *et al.* Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. *Am. J. Hum. Genet.* **100**, 605–616 (2017).
66. Firestein, G. S. Evolving concepts of rheumatoid arthritis. *Nature* **423**, 356–361 (2003).
67. Terao, C. *et al.* Genetic landscape of interactive effects ofHLA-DRB1alleles on susceptibility to ACPA(+) rheumatoid arthritis and ACPA levels in Japanese population. *J. Med. Genet.* **54**, 853–858 (2017).
68. Terao, C., Raychaudhuri, S. & Gregersen, P. K. Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis. *Annu. Rev. Genomics Hum. Genet.* **17**, 273–301 (2016).
69. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).
70. Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* **28**, 445–489 (2010).
71. Vahedi, G. *et al.* Super-enhancers delineate disease-associated regulatory nodes in T cells. *Nature* **520**, 558–562 (2015).
72. Oliveira, T. V., Maniero, F., Santos, M. H. H., Bydlowski, S. P. & Maranhão, R. C. Impact of high cholesterol intake on tissue cholesterol content and lipid transfers to high-density lipoprotein. *Nutrition* **27**, 713–718 (2011).
73. Mokhtari, R. & Lachman, H. M. The Major Histocompatibility Complex (MHC) in Schizophrenia: A Review. *J. Clin. Cell. Immunol.* **7**, (2016).
74. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).
75. Fonseka, C. Y., Rao, D. A. & Raychaudhuri, S. Leveraging blood and tissue CD4+ T cell heterogeneity at the single cell level to identify mechanisms of disease in rheumatoid arthritis. *Curr. Opin. Immunol.* **49**, 27–36 (2017).
76. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).

77. Chen, W., Mcdonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating Functional Annotations for Fine-Mapping Causal Variants in a Bayesian Framework Using Summary Statistics. (2016). doi:10.1534/genetics.116.188953

78. Koues, O. I. *et al.* Distinct Gene Regulatory Pathways for Human Innate versus Adaptive Lymphoid Cells. *Cell* **165**, 1134–1146 (2016).

79. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

80. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. doi:10.1101/445874

81. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

82. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

83. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

84. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

85. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–5, 405e1–3 (2013).

86. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

87. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).

88. Sharp, S. A. *et al.* Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis. *Diabetes Care* **42**, 200–207 (2019).

89. Kullo, I. J. *et al.* Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial). *Circulation* **133**,

1181–1188 (2016).

90. Natarajan, P. *et al.* Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091–2101 (2017).

91. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

92. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).

93. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018).

94. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

95. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).

96. Márquez-Luna, C. *et al.* Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* 375337 (2018). doi:10.1101/375337

97. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

98. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).

99. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022–1034.e6 (2019).

100. Kawakami, E., Nakaoka, S., Ohta, T. & Kitano, H. Weighted enrichment method for prediction of transcription regulators from transcriptome and global chromatin immunoprecipitation data. *Nucleic Acids Res.* **44**, 5010–5021 (2016).

101. Ishigaki, K., Akiyama, M., Kanai, M. & Takahashi, A. Large scale genome-wide association study in a

Japanese population identified 45 novel susceptibility loci for 22 diseases. *bioRxiv* (2019).

102. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

103. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).

104. Peterson, R. E. *et al.* Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell* **179**, 589–603 (2019).

105. Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

106. Gibbs, R. A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

107. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

108. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).

109. Dey, K. K., Van de Geijn, B., Kim, S. S. & Hormozdiari, F. Evaluating the informativeness of deep learning annotations for human complex diseases. *bioRxiv* (2019).

110. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep learning library for sequence data. *Nature Methods* **16**, 315–318 (2019).

111. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

112. Hirata, M. *et al.* Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *J. Epidemiol.* **27**, S9–S21 (2017).

113. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

114. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

115. Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and

interpretation of SNP-based heritability. *Nat. Genet.* **49**, 1304–1310 (2017).

116. Brown, B. C., Ye, C. J., Price, A. L., Zaitlen, N. & Asian Genetic Epidemiology Network-Type 2 Diabetes. Transethnic genetic correlation estimates from summary statistics. doi:10.1101/036657

117. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

118. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genet. Epidemiol.* **36**, 214–224 (2012).

119. Gillett, A. C., Vassos, E. & Lewis, C. M. Transforming Summary Statistics from Logistic Regression to the Liability Scale: Application to Genetic and Environmental Risk Scores. *Hum. Hered.* **83**, 210–224 (2018).

120. Bureau, S. Portal site of official statistics of Japan. (2010).

121. Mukherjee, M. *et al.* The epidemiology, healthcare and societal burden and costs of asthma in the UK and its member nations: analyses of standalone and linked national databases. *BMC Med.* **14**, 113 (2016).

122. Abhishek, A. *et al.* Rheumatoid arthritis is getting less frequent—results of a nationwide population-based cohort study. *Rheumatology* **56**, 736–744 (2017).

123. Context | Prostate cancer: diagnosis and management | Guidance | NICE.

124. Hinton, W. *et al.* Incidence and prevalence of cardiovascular disease in English primary care: a cross-sectional and follow-up study of the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC). *BMJ Open* **8**, e020282 (2018).

125. Forouhi, N. G. *et al.* Diabetes prevalence in England, 2001—estimates from an epidemiological model. *Diabet. Med.* **23**, 189–197 (2006).

126. Roadmap Epigenomics, C. *et al.* Heravi-428 Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. Integrative analysis of 111 reference human 429 epigenomes. *Nature* **518**, 317–330 (2015).

127. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

128. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese

population. *Nat. Commun.* **10**, 4393 (2019).

129. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).

130. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

131. Drake, L. Y. *et al.* B cells play key roles in th2-type airway immune responses in mice exposed to natural airborne allergens. *PLoS One* **10**, e0121660 (2015).

132. Buttari, B., Profumo, E. & Riganò, R. Crosstalk between red blood cells and the immune system and its impact on atherosclerosis. *Biomed Res. Int.* **2015**, 616834 (2015).

133. Anderson, H. L., Brodsky, I. E. & Mangalmurti, N. S. The Evolving Erythrocyte: Red Blood Cells as Modulators of Innate Immunity. *J. Immunol.* **201**, 1343–1351 (2018).

134. Lui, J. C. & Baron, J. Mechanisms limiting body growth in mammals. *Endocr. Rev.* **32**, 422–440 (2011).

135. Maier, A. B., van Heemst, D. & Westendorp, R. G. J. Relation between body height and replicative capacity of human fibroblasts in nonagenarians. *J. Gerontol. A Biol. Sci. Med. Sci.* **63**, 43–45 (2008).

136. Murphy, R. A. *et al.* Adipose tissue, muscle, and function: potential mediators of associations between body weight and mortality in older adults with type 2 diabetes. *Diabetes Care* **37**, 3213–3219 (2014).

137. Heymsfield, S. B., Gallagher, D., Mayer, L., Beetsch, J. & Pietrobelli, A. Scaling of human body composition to stature: new insights into body mass index. *Am. J. Clin. Nutr.* **86**, 82–91 (2007).

138. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).

139. Galinsky, K. J. *et al.* Estimating cross-population genetic correlations of causal effect sizes. *Genet. Epidemiol.* **43**, 180–188 (2019).

140. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *bioRxiv* 803452 (2019). doi:10.1101/803452

141. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*

47, 1236–1241 (2015).

142. Lu, Q. *et al.* A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. *Am. J. Hum. Genet.* **101**, 939–964 (2017).

143. Gusev, A. *et al.* Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nat. Commun.* **7**, 10979 (2016).

144. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

145. Bitarello, B. D. & Mathieson, I. Polygenic scores for height in admixed populations. *bioRxiv* (2020).

146. Marnetto, D. *et al.* Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat. Commun.* **11**, 1628 (2020).

147. Nathan, A., Baglaenko, Y., Fonseka, C. Y., Beynor, J. I. & Raychaudhuri, S. Multimodal single-cell approaches shed light on T cell heterogeneity. *Curr. Opin. Immunol.* **61**, 17–25 (2019).

148. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).

149. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

150. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–9 (2015).

151. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

152. Gate, R. E. *et al.* Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).

153. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **48**, 206–213 (2016).

154. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1548.e16 (2018).

155. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human

hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

156. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.* **51**, 1494–1505 (2019).

157. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309–1324.e18 (2018).

158. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature Genetics* **51**, 683–693 (2019).

159. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

160. Lal, A. *et al.* AtacWorks: A deep convolutional neural network toolkit for epigenomics. *bioRxiv* 829481 (2020). doi:10.1101/829481

161. Nathan, A. *et al.* Multimodal memory T cell profiling identifies a reduction in a polyfunctional Th17 state associated with tuberculosis progression. 2020.04.23.057828 (2020). doi:10.1101/2020.04.23.057828

162. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

163. Fang, R. *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv* 615179 (2019). doi:10.1101/615179

164. Ray, A., Khare, A., Krishnamoorthy, N., Qi, Z. & Ray, P. Regulatory T cells in many flavors control asthma. *Mucosal Immunol.* **3**, 216–229 (2010).

165. Negative Binomial Regression.

166. Weissbrod, O. *et al.* Functionally-informed fine-mapping and polygenic localization of complex trait heritability. *bioRxiv* 807792 (2019). doi:10.1101/807792

167. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. 501114 (2019). doi:10.1101/501114

168. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide

association studies. *Bioinformatics* **32**, 1493–1501 (2016).

169. Sánchez-Martín, L. *et al.* The chemokine CXCL12 regulates monocyte-macrophage differentiation and RUNX3 expression. *Blood* **117**, 88–97 (2011).

170. Gane, J. M., Stockley, R. A. & Sapey, E. TNF-α Autocrine Feedback Loops in Human Monocytes: The Pro- and Anti-Inflammatory Roles of the TNF-α Receptors Support the Concept of Selective TNFR1 Blockade In Vivo. *J Immunol Res* **2016**, 1079851 (2016).

171. Martín-Orozco, E., Norte-Muñoz, M. & Martínez-García, J. Regulatory T Cells in Allergy and Asthma. *Front Pediatr* **5**, 117 (2017).

172. Elliot, J. G. *et al.* Fatty airways: implications for obstructive disease. *Eur. Respir. J.* **54**, (2019).